Title

Classical Thermodynamics Beyond the Classical Domain

Author

Chua, Eugene Yew Siang

Publication Date

2023

UNIVERSITY OF CALIFORNIA SAN DIEGO

Classical Thermodynamics Beyond the Classical Domain

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Philosophy

by

Eugene Yew Siang Chua

Committee in charge:

Professor Craig Callender, Chair
Professor Eddy Keming Chen
Professor Brian Keating
Professor Kerry McKenzie

2023

The Dissertation of Eugene Yew Siang Chua is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my late father, who would almost have made it for my graduation,

to my mother, who always did everything she could for me,

to my siblings, who were (almost) always understanding of my choices,

and to my partner, for her unending encouragement.

EPIGRAPH

A theory is the more impressive the greater the simplicity of its premises, the more different kinds of things it relates, and the more extended its area of applicability.

Therefore the deep impression that classical thermodynamics made upon me.

It is the only physical theory of universal content which I am convinced will never be overthrown, within the framework of applicability of its basic concepts.

Albert Einstein

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I want to thank my committee members, Craig Callender, Eddy Keming Chen, and Kerry McKenzie, for their unfaltering encouragement and kindness throughout my doctoral studies, without whom my PhD research would not have been possible.

I also want to thank Nick Huggett and Chris Wüthrich at the Beyond Spacetime group at University of Illinois Chicago and Université de Genève, as well as the Templeton Foundation, for supporting me as a predoctoral fellow at the group, without which a part of this dissertation would not have been possible.

I also want to thank my partner, Sai Ying Ng, for her unwavering support and her help with proof-reading my papers.

Chapter 1, "Degeneration and Entropy", is, in full, a reprint of the material as it appears in the special issue of *Kriterion – Journal of Philosophy* 36 (2), 123-155, 2022, on *Lakatos's Undone Work: The Practical Turn and the Division of Philosophy of Mathematics and Philosophy of Science* (eds. S. Nagler, H. Pilin, and D. Sarikaya). The dissertation author was the sole author of this paper.

Chapter 2, "Does von Neumann Entropy Correspond to Thermodynamic Entropy?" is, in full, a reprint of the material as it appears in *Philosophy of Science* 88 (1), 145-168, 2021. The dissertation author was the sole author of this paper.

Chapter 3, "T Falls Apart: On the Status of Classical Temperature in Relativity", is, in full, a reprint of the material as it is forthcoming in *Philosophy of Science* 90 (5), 2023. The dissertation author was the sole author of this paper.

Chapter 5, "The Time in Thermal Time", in part, is currently under review at *Philosophy of Science*. The dissertation author was the sole author of this paper.

VITA

2018–2023 Doctor of Philosophy, Philosophy, University of California San Diego

2022 Predoctoral Fellow, Beyond Spacetime Group, University of Illinois Chicago

2013–2017 Master of Arts and Bachelor of Arts (Cantab), Philosophy, Wolfson College, University of Cambridge (Double First Class Honors)

PUBLICATIONS

2023 **Chua, Eugene Y. S.** "T Falls Apart: On the Status of Classical Temperature in Relativity." *Philosophy of Science* 90 (5), forthcoming. Preprint available at: http://philsci-archive.pitt.edu/20744/

2022 **Chua, Eugene Y. S.** "Degeneration and Entropy." *Kriterion – Journal of Philosophy* 36 (2). Special Issue on Lakatos's Undone Work: The Practical Turn and the Division of Philosophy of Mathematics and Philosophy of Science (eds. S. Nagler, H. Pilin, and D. Sarikaya). https://doi.org/10.1515/krt-2021-0032.

2021 **Chua, Eugene Y. S.** and Craig Callender. "No Time for Time from No-Time." *Philosophy of Science*, volume 88(5), pp. 1172–1184. https://doi.org/10.1086/714870.

2021 **Chua, Eugene Y. S.** "Does von Neumann Entropy Correspond to Thermodynamic Entropy?" *Philosophy of Science*, volume 88(1), pp. 145–168.
https://doi.org/10.1086/710072.

2017 **Chua, Eugene Y. S.** "An Empirical Route to Logical 'Conventionalism'." In Alexandru Baltag, Jeremy Seligman, and Tomoyuki Yamada, editors, Logic, Rationality, and Interaction, pages 631–636. Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-662-55665-8_43

ABSTRACT OF THE DISSERTATION

Classical Thermodynamics Beyond the Classical Domain

by

Eugene Yew Siang Chua

Doctor of Philosophy in Philosophy

University of California San Diego, 2023

Professor Craig Callender, Chair

Physicists have historically taken the concepts of classical thermodynamics to be universally applicable, well-understood, and secure. Thus Eddington's famous proclamation that "the law that entropy always increases, holds, I think, the supreme position among the laws of Nature", and that "if your theory is found to be against the second law of thermodynamics, I can give you no hope; there is nothing for it but to collapse in deepest humiliation." (1928, 74) Somewhat more cautiously, Einstein remarked that "[classical thermodynamics] is the only physical theory of universal content which I am convinced will never be overthrown, *within the framework of applicability of its basic concepts*." (1946, 33, emphasis mine) Suffice to say, classical thermodynamics is accorded a special status few other physical theories could dream of having.

This can still be observed in contemporary physics, where classical thermodynamic concepts are typically borrowed wholesale and applied into new domains like black hole physics and quantum gravity research by physicists like Bekenstein, Hawking, and others.

My task here is to subject this faith in classical thermodynamics to philosophical scrutiny. What *is* the ‘framework of applicability’ for thermodynamic concepts, and what are its limits? In this vein, my dissertation critically examines and challenges the foundations of thermodynamics by studying the historical trajectory of classical thermodynamic concepts and their justifications, as well as the extent to which these justifications can be carried over to new domains of inquiry. In particular, I survey how classical thermodynamic concepts extend into the domains of information theory, quantum mechanics, special relativity, general relativity, and quantum gravity. I argue that the ‘framework of applicability’ of classical thermodynamic concepts is more limited than typically assumed, and more open questions remain in the foundations of thermodynamics than one might expect.

Preface

Classical thermodynamics – the science of macroscopic properties (e.g. temperature, volume, or pressure) of ordinary systems such as steam engines, boxes of gases, or cups of coffee – has been tremendously empirically successful. Due to its seemingly ‘solved’ status, physicists often assume that classical thermodynamic reasoning is applicable to other domains: very small quantum-mechanical systems, very fast special-relativistic systems, very large systems like black holes, and even exotic quantum gravity regimes.

Historically, prominent physicists have taken the concepts of classical thermodynamics to be universally applicable, well-understood, and secure. Thus Eddington’s famous proclamation that “the law that entropy always increases, holds, I think, the supreme position among the laws of Nature”, and that “if your theory is found to be against the second law of thermodynamics, I can give you no hope; there is nothing for it but to collapse in deepest humiliation.” (1928, 74) Somewhat more cautiously, Einstein remarked that “[classical thermodynamics] is the only physical theory of universal content which I am convinced will never be overthrown, *within the framework of applicability of its basic concepts*.” (1946/1979, 33, emphasis mine) Planck (1920) in his Nobel prize speech says likewise that “the main principles of thermodynamics from the classical theory will not only rule unchallenged but will more probably become correspondingly extended” even in light of the new quantum theory.

Suffice to say, classical thermodynamics is accorded a special status few other physical theories could dream of having. This can still be observed in contemporary physics as well, Classical thermodynamic concepts like entropy, temperature and pressure are typically borrowed wholesale and applied into new domains like black hole physics and quantum gravity research by physicists like Bekenstein, Hawking, and others.

Likewise, philosophers of science in the 20th century have tended to treat the concepts of classical thermodynamics in a similar fashion by borrowing them wholesale into new domains. For instance, Reichenbach (1956) applied the classical concept of entropy (and its associated gradient) to space-time in order to ground his proposed arrow of time, though, as Earman (1974, 20) notes, there are considerable difficulties in "joining traditional thermodynamics and statistical mechanics with relativity theory". Earman charitably concedes to Reichenbach, "Whatever one's decision about the Lorentz transformation properties of thermodynamic quantities, the entropy of a thermodynamic system as measured in the rest frame of that system is a meaningful notion, and this is enough for Reichenbach's purposes." However, for our purposes, we might want to ask whether relativistic thermodynamics per se is likewise meaningful in light of these 'considerable difficulties', or runs into conceptual knots. In Chapter 3, I investigate this question specifically for relativistic temperature, and show that trouble does arise.

Similarly, in discussions of reduction and emergence, 20th century philosophers of science have generally taken for granted the existence of supposedly straightforward translations between temperature and kinetic energy, thermodynamic entropy and statistical mechanical entropy, heat and energy, and so on. Ernest Nagel (e.g. 1968) famously took there to be a clear reductive relationship between the concepts of thermodynamics and statistical mechanics, using that as a paradigmatic example of his general theory of reduction. Others like Thomas Nickles (1973) argued that there is more of a complicated relationship between these concepts involving approximations and limit-taking (what he termed reduction$_2$), though he did not really delve into the nature of these approximations in his work.

However, I believe that the qualification Einstein made deserves more philosophical scrutiny: what *is* the framework of applicability of the canonical thermodynamic concepts? Is it universal? That is, are there no limits to their applicability? Hasok Chang (2004) reminds us that the establishment of the thermodynamic concept of temperature, even in the classical thermodynamic domain, was highly non-trivial. Their development involved iterative and messy interactions between theory and experiments. Much has been done by philosophers of science to study how to justify the extension of thermodynamic concepts into new domains. Lurking in the background is Callender's (2001) warning against 'taking thermodynamics too seriously': can these classical thermodynamic concepts be extended and applied to new domains without issue? Do they 'break' or 'fall apart' in these new domains, or is there a smooth transition? For instance, the works of Jeremy Butterfield (e.g. 2011), Craig Callender (e.g. 1999), John Norton (e.g. 2016), Patricia Palacios (2018), Christian Wüthrich (2017), Roman Frigg and Charlotte Werndl (e.g. 2021), among many others, explore and critique the approximations that go into the extension of thermodynamic concepts to new domains, sometimes vindicating them, and other times not. The existence of this debate emphasizes that the applicability of thermodynamic concepts is not *a priori* universal and of utmost generality, and there may well be limits to how we can extend them. There may well be limits to the 'framework of applicability' of these thermodynamic concepts.

I see my dissertation as part of this philosophical tradition. Here, I take some preliminary steps towards investigating the various conceptual moves one might take in extending classical thermodynamic concepts like temperature and entropy beyond their original use in macroscopic scenarios, via approximations, idealizations, and/or introduction of new notions into the discourse.

The question of thermodynamics' universality has implications for both theoretical physics and philosophy. After all, thermodynamics is often understood as one of the three 'pillars of physics', alongside quantum mechanics and general relativity. Correspondingly, much work in theoretical physics goes towards the application of thermodynamics in aforementioned new domains in order to search for clues to new physics, such as the postulated statistical basis of black hole physics. But the foundations of the thermodynamic pillar might be shakier than one would like. To briefly mention one worry at the intersection of both: there's a question about where the universe's direction of time comes from. Philosophers and physicists alike have tried to explain the universe's directionality through thermodynamics: entropy represents the irreversibility of processes; this quantity always increases over time (e.g. Albert 2000). Elsewhere, physicists have studied black holes – big and enormous structures in space made of pure gravitational energy – in terms of thermodynamics (e.g. Bekenstein 1973). Both of these explanations relying on thermodynamics require that we treat very foreign things, the entire universe, or huge black holes, just as we would familiar things like a cup of hot coffee or a glass of cold beer. That is, we are assuming the universality of the applicability of thermodynamic concepts, as well as explanations couched in these concepts. But does this stretch thermodynamics too far?

The essays here inaugurate my investigation of this question by exploring and evaluating how some of these classical thermodynamic concepts were extended and generalized beyond the classical domain, into five new domains: information theory, quantum mechanics, special relativity, general relativity, and quantum gravity. In most of these domains, it turns out that the application of classical thermodynamics might not be as straightforward as we'd like it to be. That is, the framework of applicability might not be as universal as we'd like to think.

Laying out the themes of the dissertation more explicitly, Chapter 1, "Degeneration and Entropy" provides an exegesis of Lakatos's Proofs and Refutations (1976) and excavates a general account for how concepts might degenerate – by exhibiting authoritarianism and superfluity – as they get extended and stretched. I take this account to be complementary to Lakatos's account of growth and degeneration in the Methodology of Scientific Research Programmes (1978). This account is then applied to the development of the statistical mechanical concept of entropy and its extension into information theory. Specifically, it focuses on the progression of entropy as a thermodynamic concept to one laden with information-theoretic notions, notably Jaynes' arguments in his landmark 1957 paper, which I conclude to be a degenerate transition.

The themes in Chapter 1 echo through the remainder of the chapters. In the quantum mechanical domain, Chapter 2, "Does von Neumann Entropy Correspond to Thermodynamic Entropy?", evaluates the extension of the concept of entropy from another angle, by re-examining the argument that the quantum-mechanical von Neumann entropy is genuinely thermodynamic. That is, it evaluates the extension of the concept of entropy from the classical thermodynamic domain into the quantum mechanical one. Conventional wisdom holds this extension to be unproblematic, but Hemmo & Shenker (2006) argues against this view by attacking von Neumann's (1955) landmark argument. I argue that Hemmo & Shenker's arguments fail due to several misunderstandings: about statistical mechanical and thermodynamic domains of applicability, about the nature of mixed states, and about the role of approximations in physics. As a result, their arguments fail in all the cases they discussed: in the single-particle case, the finite-particles case, and the infinite-particles case.

In the relativistic realm, Chapter 3, "T Falls Apart: On the Status of Classical Temperature in Relativity", evaluates the extension of the classical concept of temperature into special relativity. Given widespread use of notions like "temperature" in e.g. the relativistic domain of black hole physics, one might expect there to be an unproblematic notion of relativistic temperature. However, I assess the numerous attempts to extend classical temperature into the special relativistic domain and argue that it 'falls apart' in a specific sense. I examine four consilient procedures for establishing the classical temperature: the Carnot process, the thermometer, kinetic theory, and black-body radiation. I argue that their relativistic counterparts demonstrate no such consilience in defining the relativistic temperature. Hence, classical temperature doesn't appear to survive a relativistic extension. I suggest two interpretations for this situation: eliminativism akin to simultaneity, or pluralism akin to rotation.

In the general relativistic realm, Chapter 4, "Do Black Holes Evaporate? The Case of Quasi-Stationarity", argues against the extension of thermodynamics into relativistic black hole physics by arguing against a typical argument for black hole evaporation. Since Hawking first predicted that black holes lose mass and 'evaporate' via Hawking radiation, the phenomenon has become a linchpin of black hole research and a key motivation for taking black holes to be thermodynamic. However, I argue against a typical derivation of black hole evaporation which essentially relies on the use of the idealization of quasi-stationarity. I argue that this idealization cannot be suitably de-idealized, and hence cannot yet be interpreted realistically. This being the case, these idealizations are not justified for application to our real world, and hence cannot support the argument for the evaporation of realistic black holes yet. Hence, the extension of thermodynamics to black holes might be less secure than previously envisioned.

Finally, Chapter 5, "The Time in Thermal Time", examines and critiques Connes & Rovelli's (1994) thermal time hypothesis and its attempt to extend and apply the concept of thermal equilibrium into quantum gravity research. Attempts to quantize gravity in the Hamiltonian approach lead to the 'problem of time'; the resultant formalism is often said to be 'frozen' and fundamentally timeless. The thermal time hypothesis responds by proposing that time emerges thermodynamically from a fundamentally timeless ontology. We define time in terms of thermal statistical states against the background of a certain algebraic structure, the C^*-algebraic structure. These statistical states define a time according to which they are in equilibrium. To avoid circularity, however, we had better have a grasp on notions like 'equilibrium' and the algebraic structure, independent of time. However, I argue that these concepts implicitly presuppose some notion of time, and cannot be extended justifiably yet to the fundamentally timeless context.

I acknowledge that my analysis here is by no means exhaustive. However, I believe it paves the path forward for a much larger research program, one that continues to engage with the question of the limits of classical thermodynamics. This will involve the careful study of other thermodynamic concepts and how they are extended into various domains.

There are many more loose ends which I plan to investigate in the future. For instance, while one might take the approach to equilibrium to be on rock-solid foundations, it turns out that there are open questions surrounding the status of equilibration in quantum mechanics and the so-called Eigenstate Thermalization Hypothesis.

In relativistic physics, there are also many open questions which can benefit from philosophical scrutiny and rumination. This dissertation only examines the relativistic temperature, but deeper investigation reveals that almost all of the canonical thermodynamic concepts rest on potentially shaky grounds.

Firstly, the status of equilibrium and the Zeroth law of thermodynamics in light of the Tolman effect in general relativity deserves more attention, for these revise the very concepts of temperature, heat flow, and equilibrium, that we are used to in classical thermodynamics: the equivalence of the measured temperature of two systems no longer entails no heat flow, and their *non*-equivalence does not entail heat flow either, because of general relativistic effects. I hope to work out the philosophical and conceptual implications of this effect in the future.

Secondly, the concept of entropy is often assumed to be Lorentz-invariant, owing to old arguments by Planck (1908) among others. However, it's been noted by Gavassino (2021) that this argument is circular: it assumes the possibility of reversible Lorentz boosts, but this amounts to assuming the invariance of entropy under Lorentz boosts. I believe it remains an open question whether there is a watertight argument for the Lorentz-invariance of entropy.

Likewise, the concept of pressure is also assumed to be Lorentz-invariant. Standard arguments, however, tend to claim that the pressure simply *is* the rest-frame pressure, rather than show that the pressure is Lorentz-invariant. Sutcliffe (1965) furthermore disputes the Lorentz-invariance of pressure by disambiguating thermodynamic and mechanical pressure, arguing that the rest-frame thermodynamic pressure p_0 transforms to the moving-frame thermodynamic pressure p' via $p' = \gamma^2 p_0$ instead (where γ is the Lorentz factor). In the future, I hope to inspect these arguments in further detail.

Finally, even the concept of a force has been placed under scrutiny. As Landsberg & Johns (1970) point out, differences in two proposed relativistic temperature transformations – whether a moving body is cooler or hotter – can be traced back to a difference in two choices of the force law, both of which coincide in the rest frame. Is there a genuine dispute here, and to what extent can this worry be settled?

It thus seems that classical thermodynamics, despite its enormous empirical success in the rest frame and in the macroscopic world, rests on shaky foundations when extended to other domains. Only further work can reveal the extent to which this shakiness should be worrying. Unfortunately, these loose ends deserve more care and rumination than is allowed by the present space and time constraints. I thus leave them for future work.

Eugene Y. S. Chua

University of California San Diego

June 2023

Bibliography

Butterfield, J. (2010). "Less is different: Emergence and reduction reconciled". In: *Foundations of Physics* 41.6, pp. 1065–1135. DOI: 10.1007/s10701-010-9516-1.

Callender, C. (2001). "Taking thermodynamics too seriously. Studies in History and Philosophy of Science Part B". In: *Studies in History and Philosophy of Modern Physics* 32.4, pp. 539–553. DOI: 10.1016/s1355-2198(01)00025-9.

Chang, Hasok (2004). *Inventing Temperature: Measurement and Scientific Progress.* New York, US: OUP.

Earman, J. (1974). "An attempt to add a little direction to "the problem of the direction of Time"". In: *Philosophy of Science* 41.1, pp. 15–47. DOI: 10.1086/288568.

Eddington, A.S. (1928). *The nature of the physical world: Gifford Lectures (1927.* en. The University Press.

Einstein, A. (1979). *Autobiographical Notes.* en. Trans. by P.A. Schilpp. Open Court Printing.

Gavassino, L. (Dec. 2021). "Proving the Lorentz Invariance of the Entropy and the Covariance of Thermodynamics". In: *Foundations of Physics* 52.1, p. 11. ISSN: 1572-9516. DOI: 10.1007/s10701-021-00518-w. URL: https://doi.org/10.1007/s10701-021-00518-w.

Landsberg, P.T and K.A Johns (1970). "The Lorentz transformation of heat and work". In: *Annals of Physics* 56.2, pp. 299–318. ISSN: 0003-4916. DOI: https://doi.org/10.1016/0003-4916(70)90020-5. URL: https://www.sciencedirect.com/science/article/pii/0003491670900205.

Nagel, E. (1968). *The structure of science: Problems in the logic of scientific explanation.* Routledge & Kegan Paul Ltd.

Nickles, T. (1973). “Two concepts of intertheoretic reduction”. In: *The Journal of Philosophy* 70.7, pp. 181–201. DOI: 10.2307/2024906.

Norton, J.D. (2016). “The impossible process: Thermodynamic reversibility.” In: *Studies in History and Philosophy of Modern Physics* 55, pp. 43–61. DOI: 10.1016/j.shpsb.2016.08.001.

Palacios, P. (2018). “Had we but world enough, and time... but we don’t!: Justifying the thermodynamic and infinite-time limits in Statistical Mechanics”. In: *Foundations of Physics* 48.5, pp. 526–541. DOI: 10.1007/s10701-018-0165-0.

Planck, M. (1908). “On the dynamics of moving systems”. In: *Ann. Phys. Leipz.* 26, pp. 1–34.

— (1920). “The Genesis and Present State of Development of the Quantum Theory”. In: *Nobel Lecture, June 2, 1920.* URL: https://www.nobelprize.org/prizes/physics/1918/planck/lecture/.

Reichenbach, H. (1956). *The Direction of Time.* University of California Press.

Sutcliffe, W. G. (Sept. 1965). “Lorentz transformations of thermodynamic quantities”. In: *Il Nuovo Cimento (1955-1965)* 39.2, pp. 683–686. DOI: 10.1007/BF02735833. URL: https://doi.org/10.1007/BF02735833.

Wuthrich, C. (2019). “Are Black Holes About Information?” In: *Why trust a theory?: Epistemology of fundamental physics. essay.* Ed. by R. Dardashti, R. Dawid, and K. Thebault. Cambridge University Press.

Chapter 1

Degeneration and Entropy

"As a young man I tried to read thermodynamics, but I always came up against entropy as a brick wall that stopped my further progress. I found the ordinary mathematical explanation, of course, but no sort of physical idea underlying it. No author seemed even to try to give any physical idea. Having in those days great respect for textbooks, I concluded that the physical meaning must be so obvious that it needs no explanation, and that I was especially stupid on the particular subject."

– James Swinburne (1904, p. 3)

"My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name. In the second place, and more importantly, no one knows what entropy really is, so in a debate you will always have the advantage."

– Claude Shannon,
according to McIrvine and Tribus (1971, p. 180)

1.1 Introduction

Lakatos (1976/2015) argued in *Proofs and Refutations* (**P&R**) that the comprehension of mathematical concepts must be accompanied with a clear understanding of how and why these concepts came into existence. For Lakatos, a concept is to be understood in terms of a temporally extended process through which the initial, primitive, concept is continually refined. To rip a concept apart from its context of discovery or its problem-situation – the problems or questions which led to the concept's genesis and evolution – is to miss a complete understanding of it.

All of the above – concerning how to *comprehend* a concept – has been much discussed over the last few decades.[1] Less discussed is how to evaluate a concept according to the heuristic approach presented in **P&R**: how do we know whether a concept is problematic, needs rehabilitation, or, worse still, must be abandoned, given the heuristic approach? In short, how do we know whether a concept is degenerating?

That not much has been said about this is curious, since Lakatos clearly had some such standard in mind. In this paper, I explore Lakatos' views on degeneration in **P&R**, which has often been neglected for the sort of degeneration Lakatos (1978) discussed in *Methodology of Scientific Research Programmes* (**MSRP**). It seems to me that **P&R** offers new criteria for degeneration that sheds light on Lakatos's approach.

The primary goal here is to motivate an account of degeneration based on my reading of **P&R**. I propose two criteria for degeneration: superfluity, or generalization for generalization's sake, which involves the introduction of trivial extensions or terminology to a theory or concept; and authoritarianism, the introduction and employment of a concept into the discussion without justification while ignoring the problem-situation of said concept. In my view, these notions of degeneration depart from the two conditions found in **MSRP**: the former relate not to the content of a concept or theory,[2] but to their *methodological* aspects (i.e. what I'll call *depth*).

[1] See e.g. Corfield (1997), Leng (2002), and Werndl (2009).

[2] It seems to me that Lakatos did not really have a clear-cut distinction between concept and theory. For

As such, by considering the notions of degeneration in both **P&R** and **MSRP**, I propose an extended account of Lakatosian degeneration which evaluates both content and depth.

The secondary goal is to apply this extended account of degeneration to entropy. Why entropy? The concept of entropy has a tumultuous past, with many interpretations,[3] complications,[4] and disagreements colouring its rich history. This is coupled, however, with its extensive usage in countless sciences, be it in black hole thermodynamics,[5] quantum theory,[6] AdS/CFT research in theoretical physics,[7] and even biology and neuroscience.[8] There remains much room to evaluate these applications and extensions of the concept of entropy, including whether we *should* extend it in these ways.[9] This makes entropy an interesting case study for degeneration – after all, my proposed account of degeneration is intended to evaluate the extension of a concept. Since the assessment of the concept of entropy remains pretty much an open question, it is hoped that my account will provide some heuristics for beginning this assessment. Here, as proof of concept, I focus on one key transition point for the concept of entropy – the transition from thermodynamic to information-theoretic interpretations of entropy. In Lakatosian style, I identify a key piece of writing in this transition and evaluate it with the twin criteria developed here. By critiquing Jaynes's (1957) landmark paper on thermodynamics and information, I argue that this transition suffered from superfluity and authoritarianism, and hence, degeneration.

instance, he talks about the 'theory of polyhedra' throughout **P&R**, but also refers to the "concept of polyhedron" as well. I see no cost to this lack of distinction here. Hence, I follow Lakatos in using "concept" and "theory" interchangeably in this paper whenever the context is clear. Thanks to Craig Callender and Kerry McKenzie for independently raising this concern.

[3] See Uffink (2001) for how interpreting the original thermodynamic concept of entropy is itself a challenge.

[4] See e.g. Callender (1999), Chua (2021), or Goldstein et al. (2020) for some of these complications concerning how different entropies relate to others.

[5] See e.g. Dougherty & Callender (ms) or Wallace (2020).

[6] See e.g. Bub (2005).

[7] See e.g. Natsuume (2015).

[8] See Friston & Stephan (2007).

[9] See e.g. Roach (2020) who evaluates the use of entropy in biology.

1.2 Degeneration in P&R

P&R takes the form of a dialogue between a fictional teacher and his students which includes Gamma (and later Alpha, Rho and Zeta, among others). Lakatos's main target was a 'deductivist' approach to mathematics:

> This style starts with a painstakingly stated list of axioms, lemmas and/or definitions. The axioms and definitions frequently look artificial and mystifyingly complicated. One is never told how these complications arose. The list of axioms and definitions is followed by the carefully worded theorems. These are loaded with heavy-going conditions; it seems impossible that anyone should ever have guessed them. The theorem is followed by the proof. (1976/2015, p. 151)

In Lakatos's view, this approach to mathematics is misguided. By rationally reconstructing the historical development of the Euler characteristic:[10]

$$V - E + F = 2 \tag{1.1}$$

he showed that these definitions, axioms, theorems and so on developed only as the result of a long history of proofs and refutations: they are *proof-generated concepts*.

For Lakatos, actual mathematics is not deductivist. Nevertheless, it is a rational affair representative of what he called the *heuristic approach*:

> [...] deductivist style tears the proof-generated definitions off their 'proof-ancestors', presents them out of the blue, in an artificial and authoritarian way. It hides the global counterexamples which led to their discovery. Heuristic style on the contrary highlights these factors. It emphasises the problem-situation: it emphasises the 'logic' which gave birth to the new concept. (1976/2015, p. 153)

More generally, a concept is only appropriately understood when we understand its historical trajectory. In Lakatos's view, "there is a simple pattern of mathematical discovery– or of the growth of informal mathematical theories" (1976/2015, pp. 135–136), which is given by the following seven stages:

[10] The Euler characteristic was initially postulated to universally describe polyhedra in terms of their number of faces (F), vertices (V), and edges (E).

1. Primitive conjecture.
2. Proof (a rough thought-experiment or argument, decomposing the primitive conjecture into subconjectures or lemmas).
3. ‘Global’ counterexamples (counterexamples to the primitive conjecture) emerge.
4. Proof re-examined.
5. Proofs of other theorems are examined to see if the newly found lemma or the new proof-generated concept occurs in them.
6. The hitherto accepted consequences of the original and now refuted conjecture are checked.
7. Counterexamples are turned into new examples – new fields of inquiry open up.

Call this the *heuristic process*. This tells us how a mathematical theory or concept ought to grow, from a rough primitive conjecture to a proof-generated concept (and beyond), through the use of heuristics like employing counterexamples, discovering hidden lemmas, and so on. We appropriately comprehend a concept only when we start with the primitive conjecture – the genesis of the concept – and grasp the ensuing adjustments and responses to the concept through which it is precisified and stretched.

But as Gamma points out, *growth* is opposed to *degeneration*. (1976/2015, p. 103) However, while Lakatos paints a clear picture as to how mathematical theories grow, he is much less explicit about degeneration. This is curious because the term ‘degeneration’ seems to bear significant normative weight in Lakatos’s appraisal of research methodologies. My goal here is to remedy this situation by explicating Lakatos’s account of degeneration in terms of two distinct criteria.

1.2.1 Superfluity

The first criterion for degeneration appears when Alpha (1976/2015, p. 86) charts the development of the dialogue in **P&R** thus far, in response to ever-exotic counterexamples:[11]

1. One vertex is one vertex

2. $V = E$ for all perfect polygons

3. $V - E + F = 1$ for all normal open polygonal systems

4. $V - E + F = 2$ for all normal closed polygonal systems, i.e. polyhedra

5. $V - E + F = 2 - 2(n-1) + \sum_k^F e_k$ for normal n-spheroid polyhedra

6. $V - E + F = \sum_{j=1}^F 2 - 2(n_j - 1) + \sum_{k=1}^F e_{kj}$ for normal n-spheroid polyhedra with multiply-connected faces and with cavities

Alpha proclaims that "this is a miraculous unfolding of the hidden riches of the trivial starting point [(1)]". However, Rho retorts: "Hidden 'riches'? The last two only show how cheap generalisations may become!" Gamma concurs:

> (6) and (7) are not growth, but *degeneration*! Instead of going on to (6) and (7), I would rather find and explain some exciting new counterexample [to $V - E + F = 2$]! (1976/2015, p. 103, emphasis mine)

Degeneration is tied, in part, to 'cheap generalisation' in the process of developing a concept or theory. It plays an evaluative role insofar as it tells us when, in something like the chain of generalizations from (1) to (7), we should stop in response to a new counterexample. Gamma describes (7) as a pointless generalization: who cares about polyhedrons with cavities or multiply-connected faces? To Gamma, "it serves only for making up complicated, pretentious formulas for nothing." (1976/2015, pp. 102–103)

A cheap generalization, in Gamma's view, is a *trivial extension* of a concept. (1976/2015, p. 102) The generalization was not well-motivated, and thus nothing deep was gleaned from doing

[11] See Lakatos (1976) for discussion.

so, even though our formula does become more general. It is, to put it bluntly, *generalization for the sake of generalization.* When a concept is extended in order to encompass new cases, but those cases were not relevant to the original problem-situation (or the heuristics that followed), Gamma would consider them trivial extensions.

Lakatos, channelling Pólya, assents to this in his description of what cheap generalization amounts to:

> Pólya points out that shallow, cheap, generalisation is 'more fashionable nowadays than it was formerly. *It dilutes a little idea with a big terminology. The author usually prefers to take even that little idea from somebody else, refrains from adding any original observation, and avoids solving any problem except a few problems arising from the difficulties of his own terminology.*
>
> Another of the greatest mathematicians of our century, John von Neumann, also warned against this 'danger of degeneration', but thought it would not be so bad 'if the discipline is under the influence of men with an exceptionally well-developed taste'. One wonders, though, whether the 'influence of men with an exceptionally well-developed taste' will be enough to save mathematics in our 'publish or perish' age. (1976/2015, p. 104, fn. 160, emphasis mine)

For our purposes, it certainly helps that Lakatos directly connects 'shallow, cheap, generalisation' to the 'danger of degeneration'. Alternatively stated, the generalization of a concept to new domains without justification is *superfluous* and only adds unnecessary terminology ("pretentious formulas"). This, in Lakatosian terms, is concept-stretching 'gone wrong' – a concept is stretched too far without justification, resulting in trivial generalizations. This leads to the concept's degeneration. Call this first criterion of degeneration *superfluity.*

What counts as justification and can avoid superfluity? Of course, the justification must be relevant to the problem-situation at hand,[12] but what else? We can distil further insights from Lakatos's discussion of the Euler characteristic: in that case, the problem-situation is one where the key concern is about classifying (what we intuitively might call) polyhedra. However, while cavities were helpful in categorizing polyhedra based on *Definition 1*, i.e. the naïve notion

[12] Some might argue that this is too restrictive – what if a proposal has relevance elsewhere? Then it must prove its relevance to that problem-situation.

of a polyhedron as a solid, a single 'polyhedron with a cavity' corresponds to an entire class of polyhedra in the succeeding proof-generated concept of polyhedra as connected surfaces (Lakatos calls this *Definition 2*). (1976/2015, p. 16) On this definition, a polyhedron is not a *solid.* Given this background, the number of cavities does not actually pick out a unique type of polyhedra,[13] and hence any extension of the Euler characteristic which takes into account the number of cavities simply plays no real role in advancing the research of polyhedra.[14] Instead, we have merely added superfluous terminology which do not concern objects of interest.

From this we might infer that the sort of justification which vindicates any particular generalization or introduction of terminology is one that can motivate why this terminology or generalization was introduced – particularly in relation to the sort of objects or concepts we care about – and how it can possibly lead to the growth of the theory or concept. It cannot be just a *trivial pun*, which is essentially what a 'polyhedron with a cavity' is, since it is not a polyhedron per se at all given the problem-situation where Definition 2 is accepted as a proof-generated concept – it must serve some *use* and justify its own existence, so to speak.

Of course, this requirement is not precise, for it is not always clear what is trivial. Alpha questions: "You may be right after all. But who decides where to stop? Depth is only a matter of taste." (1976/2015, p. 103) In response, Gamma proposes:

> Why not have mathematical critics just as you have literary critics, to develop mathematical taste by public criticism? We may even stem the tide of pretentious trivialities in mathematical literature. (1976/2015, p. 104)

Unfortunately, Lakatos does not say much more about what 'taste' amounts to, though his comment on von Neumann's comment suggests skepticism towards its role.

Nevertheless, hindsight helps, as in the case of cavities failing to pick out the relevant sort of properties for polyhedra: 'relevance' is determined as a matter of practice, adoption,

[13] See Lakatos (1976/2015, p. 97, fn. 150(a)).

[14] Likewise for multiply-connected faces: it was simply not interesting for research because our means of classifying polyhedra simply did not require multiply-connected faces. As Lakatos notes, as the theory of polyhedra evolved, "the problem of how multiply-connected faces influence the Euler-characteristic of a polyhedron lost all its interest" (1976/2015, p. 97, fn. 150(c)).

and actual contribution to the problem-situation, e.g. *Definition 2* as an improvement over *Definition 1*. As Kiss puts it succinctly (though in the context of the **MSRP**): "One step in a research program can be treated as progressive or degenerating only in hindsight, when we see future developments. Appraisal of research programs is as fallible as the theories themselves." (2006, p. 316)

We can do better in other cases. In line with the heuristic approach, conscious awareness of the problem-situation – and the heuristics associated with it – allows one to avoid superfluity. Lakatos discusses the example of the mathematician Becker, who aimed to provide a conclusion to the classification problem by providing a new generalization to the Euler characteristic:

$$V - E + F = 4 - 2n + q \tag{1.2}$$

where n is the number of cuts that is needed to divide the polyhedral surface into simply-connected surfaces for which $V - E + F = 1$, and q is the number of diagonals that one has to add to reduce all faces to simply-connected ones. (1976/2015, p. 103, fn. 158) Unfortunately for Becker, Lakatos notes that the mathematicians Lhuilier and Jordan had already written about this over half a century ago, except in different terminologies. While Becker's work does count as a valid generalization of Euler's original formulation, it was ultimately trivial – adding only new terminology – and did not contribute to the development of the concept. Here it is clear that a cognizance of the problem-situation would have helped: being aware of the concept's past iterations, problems, and errors, one learns what not to do.

By being aware of the problem-situation for the concept, we can avoid trivial generalizations. Beyond that, we can only hope to pinpoint degeneration in retrospect unless we are blessed with the great gift of 'exceptional taste' in the subject matter at hand (in which case, a certain sort of clairvoyance – of what *will* work out in research – is possible). However, given Lakatos's distaste for formalism and inclination towards informal heuristics – which are themselves not always precise – this might be to Lakatos's liking.

1.2.2 Authoritarianism

In discussing the heuristic approach, Lakatos notes a common response to byzantine definitions (in this case, Carathéodory's definition of a measurable set) with disapproval:

> Of course there is always the easy way out: mathematicians define their concepts just as they like. But *serious teachers do not take this easy refuge.* Nor can they just say that this is the correct, true definition … and that mature mathematical insight should see it as such. (1976/2015, p. 162)

There is something problematic with the deductivist style of simply introducing concepts out of thin air without appropriately situating them within the proof's problem-background. This approach is *authoritarian*:

> One can easily give more examples, where stating the primitive conjecture, showing the proof, the counterexamples, and following the heuristic order up to the theorem and to the proof-generated definition would dispel the *authoritarian mysticism of abstract mathematics, and would act as a brake on degeneration.* A couple of case-studies in this degeneration would do much good for mathematics. Unfortunately the deductivist style and the atomisation of mathematical knowledge protect '*degenerate*' papers to a very considerable degree. (1976/2015, p. 163, emphasis mine)

This provides a new criterion for degeneration, which occurs when a concept (or terminology, or definition, or perhaps even a theory) is introduced without justification into some line of inquiry. New concepts are instead used with the attitude "that this is the correct, true definition" without qualification. This adds a 'mystical' and 'authoritarian' element to these new concepts which ignores the background problem-situation leading to that line of inquiry to begin with. Call this criterion for degeneration *authoritarianism.*

While many of Lakatos's examples of authoritarian methodology were textbooks (like Rudin or Halmos), I should emphasize that there is no reason to interpret his discussions about authoritarianism as something which only applied to pedagogical works.[15] For instance, in the quote above, Lakatos was clearly discussing how the dispelling of "authoritarian mysticism" is needed to remove the protection of degenerate papers, i.e. research, by the deductivist style of

[15] I thank an anonymous reviewer for pushing me to clarify this.

mathematics. Furthermore, in Appendix I, Lakatos writes that:

> It was the infallibilist philosophical background of Euclidean method that bred the authoritarian traditional patterns in *mathematics*, that prevented publication and discussion of conjectures, that made impossible the rise of mathematical criticism. (1976/2015, p. 147, emphasis mine)

It was the infallibilist philosophical background of Euclidean method that bred the authoritarian traditional patterns in mathematics, that prevented publication and discussion of conjectures, that made impossible the rise of mathematical criticism. (1976/2015, p. 147, emphasis mine)

Lakatos is here using authoritarianism as a criterion for evaluating mathematics *as a whole*, including research like publications and discussions, rather than pedagogy in particular. As such, I believe that authoritarianism should be interpreted as a criterion of evaluating degeneration which is applicable to research as well as pedagogy.

Lakatos raises the example of Rudin's discussion of bounded variation. While introducing the Riemann-Stieltjes integral, Rudin introduces the notion of bounded variation.[16] He then proves a theorem to the effect that a function of bounded variation, satisfying other criteria, is also a member of the class of Riemann-Stieljes integrable functions. However, Lakatos accuses Rudin of failing to explain why the Riemann-Stieljes integral and bounded variation were relevant to begin with:

> So now we have got a theorem in which two mystical concepts, bounded variation and Riemann-integrability, occur. But two mysteries do not add up to understanding. Or perhaps they do for those who have the 'ability and inclination to pursue an abstract train of thought'? (1976/2015, p. 156)

The *non-degenerative* way of presenting the concept would have shown that the two concepts arose as proof-generated concepts out of the same problem-situation:

> A heuristic presentation would show that both concepts – Riemann-Stieltjes integrability and bounded variation – are proof-generated concepts, originating in one and the same proof: Dirichlet's proof of the Fourier conjecture. This proof gives the problem-background of both concepts. (1976/2015, p. 156)

Lakatos notes that Rudin does mention this, but in a way disconnected from the two aforemen-

[16] For Lakatos's discussion, see Lakatos (1976/2015, pp. 155–162).

tioned concepts: it was hidden in some exercise in a different chapter. Lakatos declares that the two concepts, introduced this way, were introduced in "an authoritarian way" (1976/2015, p. 156, fn. 12).[17]

Thus, both superfluity and authoritarianism arise from a failure to grapple with the problem-situation. While superfluity arises from trivial generalizations of concepts arising from this lack of awareness, authoritarianism arises from the unmotivated introduction and application of concepts.

For Lakatos, to 'introduce' a term out of the blue into a discussion is a "magical operation which is resorted to very often in history written in deductivist style!" (1976/2015, 157, fn. 17) To treat bounded variation or the Riemann-Stieltjes integral as an "introduced" concept rather than a proof-generated one – as he shows in (1976/2015, pp. 156–162) – is to miss out on the understanding of the concept. On the heuristic approach,

> [. . .] the two mysterious definitions of bounded variation and of the Riemann-integral are *entzaubert*, deprived of their authoritarian magic; their origin can be traced to some clear-cut problem situation and to the criticism of previous attempted solutions of these problems. (1976/2015, p. 158)

In sum, I understand authoritarianism – the second criterion of degeneration – as introducing new concepts into some line of inquiry without justification, while ignoring the problem-situation and the heuristics that led us to that discourse to begin with.

1.2.3 Their Normative Import

Recall that the heuristic approach places emphasis on understanding the background and historical trajectory of a concept, over and above the concept in its current form. The original problem and the errors that followed (i.e. the problem-situation) are just as important as the end-product and the proof-generated concept, because they tell us what has not worked

[17] Some might worry that having a complete grasp or presentation on any problem-situation might be too demanding. However, Lakatos's reply to this "pedestrian argument" is: "let us try." (1976/2015, p. 153) He thinks of this stringency as a positive point: while this makes academic work demanding and long, ". . . it has to be admitted that they would be much fewer too, as the statement of the problem-situation would too obviously display the pointlessness of quite a few of them." (1976/2015, p. 153, fn. 5) Thanks to Nuhu Osman Attah for raising the worry.

(and will not worked), why the concept is the way it is now, and hopefully the available routes for development based on those errors (at least, it rules out the unavailable routes for development).

By failing to grasp this problem-situation, superfluity and authoritarianism miss out on a complete understanding of the concept *as part of a historical trajectory*, instead atomizing it as a stand-alone concept – methodologies with superfluity and authoritarian tendencies thus hinder an understanding of the concept. As Zeta puts it, "A problem never comes out of the blue. It is always related to our background knowledge". (1976/2015, p. 74) By adopting a methodology which elides this background knowledge, our understanding of a concept and the associated problem is rendered incomplete.

On one hand, authoritarianism fails to account for past problems and errors by tearing apart present discussions from past problems – the discussion is presented without context; the reader is told to accept it on faith, or that they need "mathematical maturity" to understand it. (1976/2015, p. 151, fn. 1) This obscures the errors and problems crucial to generating the proof-generated concept. We then lose sight of the direction the concept was taking in the long chain of proofs and refutations: we are "rewriting history to purge it from error" (1976/2015, p. 49) and hence "the zig-zag of discovery cannot be discerned in the end-product". (1976/2015, p. 44) The degenerate concept becomes atomized as a result, which hinders growth in Lakatos's view.

On the other hand, superfluity reflects a lack of concern for a concept's problem-situation. Terminology is produced, but not because it presents an insightful development for the concept and its trajectory. Sometimes, as in Becker's case, one mistakes themselves to be presenting fruitful development for a concept. Again, this approach treats the concept at hand as one that is divorced from its problem-situation – instead of considering what problems the terminology is meant to resolve, we are instead pursuing 'cheap, shallow generalisations' even if the result is ultimately trivial. We have seen this in the case of generalizing the Euler characteristic to (6) and (7): obviously, we can generalize the Euler characteristic to the case of intuitive polyhedra with cavities and multiply-connected faces, but the naïve terminology – of cavities and multiply-connected faces – and the accompanying generalizations were simply no longer fruitful to the discussion of polyhedra at that point of the trajectory of the concept in the dialogue. A superfluous extension of a concept involves ever more esoteric 'generalizations' to cover cases no one cares about (in the context of the line of inquiry surrounding that concept). In Lakatos's words:

> Quite a few mathematicians cannot distinguish the trivial from the non-trivial. This is especially awkward when a lack of feeling for relevance is coupled with the illusion that one can construct a perfectly complete formula that covers all conceivable cases. Such mathematicians may work for years on the 'ultimate' generalisation of a formula, and end up by extending it with a few trivial corrections. (1976/2015, p. 103, fn. 158)

By failing to grasp what is trivial (which can be aided by hindsight or a grasp of the problem-situation), research degenerates by either treading trodden grounds (as with the case of Becker) or extending a concept to domains which are simply unfruitful (as in the case of cavities and multiply-connected faces).

In short, both superfluity and authoritarianism have clear normative import: if our goal is to pursue growth for the concept by having a clearer understanding of the concept, we ought to avoid both forms of degeneration. Degeneration in these two senses thus play an evaluative role for the growth of a concept.

1.3 Degeneration in P&R vs. Degeneration in MSRP

I have focused on degeneration in the context of **P&R**, but how does this connect with Lakatos's much more famous classification of scientific research programmes as degenerative or progressive? For Lakatos in **MSRP**, scientific theories should be understood akin to the heuristic approach to mathematical theories and concepts. They are not isolated atoms but sequences of theories or concepts – what he later calls a *scientific research programme* – grouped together by various criteria (such as their positive and negative heuristics).[18] This "shifts the problem of how to appraise *theories* to the problem of how to appraise *series of theories*. Not an isolated theory, but only a series of theories can be said to be scientific or unscientific: to apply the term 'scientific' to one *single* theory is a category mistake." (24, p. 34)

The trajectory of this sequence of theories over time (i.e. its 'problemshift' from one theory to another) is progressive or degenerative according to two criteria: whether it is (i) *theoretically* progressive and (ii) *empirically* progressive. (1978, pp. 33–34) Being theoretically progressive refers to a succeeding theory containing 'excess empirical content' by predicting novel facts compared to its predecessor, and would distinguish, according to Lakatos, a 'scientific' problemshift from a non-'scientific' one. Being empirically progressive refers to the excess empirical content of this succeeding theory leading to the discovery of new facts, thereby corroborating the new theory's novel predictions. A problemshift is deemed *overall progressive* if it is *both* theoretically and empirically progressive, and *overall degenerating* if it is not.

Much ink has been spilled over this point.[19] What I am interested in is how the account of degeneration presented in **MSRP** can be augmented by the account of degeneration I have presented based on **P&R**, and how they can be collectively marshalled for the philosopher of science despite the obvious differences between the sciences and mathematics.[20]

There might be some doubt as to whether Lakatos's views on mathematics in **P&R** can be so neatly transplanted to the scientific context. Despite the differences between science and mathematics, however, Lakatos emphasized that "mathematical heuristic is very like scientific heuristic – not because both are inductive, but because both are characterised by conjectures,

[18] See Lakatos (1978, pp. 47–52).

[19] For an excellent collection of essays on Lakatos's methodology, see Kampis, Kvasz & Stöltzner (2002).

[20] Some might object that there are some differences between the sciences and mathematics that prevent such analysis. Interestingly, Hallert (1979a, 1979b) and Stöltzner (2002) did something to similar effect but in the opposite direction of what I am doing here. They do the reverse, by proposing that we can apply the application of the conditions of progress and degeneration described in MSRP to mathematics (and mathematical physics) profitably. What I am doing here can be seen as a continuation of that sort of project – to bring Lakatos's insights from his discussion of informal mathematics into physics. My analysis of Jaynes is an attempt to demonstrate that such an analysis is possible.

proofs, and refutations. The important difference lies in the nature of the respective conjectures, proofs (or, in science, explanations), and counterexamples". (1976/2015, p. 78) In my view, Lakatos would likely take the heuristic approach to be applicable to both physics (and the sciences) and mathematics.

I believe the two accounts of degeneration complement each other: while the account in **MSRP** focused on *content*, the account in **P&R** focused on *depth* – how deep or trivial the research is, how it connects with its predecessors, the potential or actual fruitfulness of the research, and so on (discussed above in II) – which in turn hinges on *methodology*.

This former account of degeneration with respect to *content* takes a central role in Lakatos's project for scientific research programmes. Furthermore, it can be straightforwardly cashed out in terms of the *number* of theoretical predictions a theory makes and actually corroborated predictions relative to its rivals and competitors. These properties make them an easy target for analysis: for starters, we can just count the number of propositions (or sentences, or whatever your favourite truth-bearers are) non-vacuously entailed by the theory (and whether they are corroborated)![21] This is not to say that considerations about content is somehow unimportant. We need yardsticks for discussing, comparing, and evaluating the content produced by scientific research programmes. If these yardsticks are clear, all the better. However, the focus on content – in terms of the predictions of a theory – does not really consider the *methodological* issues that might also be considered progressive or degenerating. I think this is important. In Rho's words in **P&R**: "not every increase in content is also an increase in depth." (1976/2015, p. 103) A theory might have superior content while simultaneously experiencing a degeneration in methodology and *depth* of research.

This is where my proposed account of degeneration comes into play. This focuses on the *depth* of research at each problemshift – and this of course depends much more on the style and methodology of research, which may also be far more diverse than a generally accepted theory and its contents. Nevertheless, just as there are authoritative interpretations of theories

[21] Of course, what counts as 'non-vacuous' may yet be a matter for contention.

even when there are generally numerous interpretations of any single theory, there are also authoritative figures, presentations, rhetoric, and methodologies, which may yet be open to analysis of depth. (An attempt to analyze one such authoritative presentation is made later in IV.) This, in turn, requires analysis of notions like taste, triviality, fruitfulness, awareness of the problem-situation and so on, as we have discussed in II, which are not obviously amenable to logical analysis like the content-oriented notions of degeneration are.

But both *are* inseparable aspects of scientific research – we need to be concerned about both the depth of research, in terms of whether authoritarianism and superfluity are occurring, and the content being produced by the research, in terms of whether there is theoretical and empirical progress.

If I am right, we can extend Lakatos's classification of scientific research programmes quite straightforwardly: a problemshift is overall progressive if it is overall progressive with respect to content (content-degeneration) and avoids degeneration with respect to depth (depth-degeneration), that is, by being theoretically and empirically progressive while avoiding authoritarianism and superfluity. It is degenerating otherwise. Depth-degeneration comes in degrees – not all research at any one time will typically contain authoritarianism and superfluity, but how much research falls afoul of authoritarianism and superfluity will determine the degree of degeneration with respect to depth. My account of degeneration, based on the **P&R**, thus augments the account of degeneration found in **MSRP** and provides a new dimension of analysis for Lakatos's overall framework.

1.4 Entropy: Degeneration from Physics to Information

As a proof of concept, I begin this project by analyzing one important shift for the concept of entropy – the incorporation of information-theoretic notions into entropy by Jaynes (1957).[22] Entropy plays a role of ever-increasing importance in our best sciences. Despite its

[22] I focus solely on the first of two papers he published on the topic in 1957, as the second part focuses largely on applying the account developed in part I, rather than presenting any novel arguments for the account itself.

ubiquity, the concept of entropy is not easily grasped. Some like Swinburne complained that "there is no sort of physical idea underlying [entropy]" (1904, p. 3); the concept of entropy does not come equipped with an obvious physical idea for us to latch onto, despite being defined in terms of physical quantities. This hints at degeneration – how can a concept that is so ubiquitous in physics be so imprecisely understood? In what follows, I examine the concept of entropy with Lakatos's method in the Appendix of **P&R**: highlight a piece of work which significantly influenced a concept's trajectory and point out its various degenerative traits, while suggesting what could have been done otherwise.

1.4.1 Jaynes's 'Information Theory and Statistical Mechanics'

Jaynes was not the first to propose a marriage between information theory and statistical mechanics – that honour goes to Leon Brillouin.[23] However, Jaynes's paper is one of the (if not the) most influential. As Seidenfeld (1986, p. 468) notes, "I doubt there is a more staunch defender of the generality of entropy as a basis for quantifying (probabilistic) uncertainty than the physicist E. T. Jaynes." In a footnote to his famous 1973 paper on black hole entropy, Bekenstein observed that "the derivation of statistical mechanics from information theory was first carried out by E. T. Jaynes". (1973, fn. 17) In that same paper, he notes that by 1973, "the connection between entropy and information is well known", (1973, p. 2335) and later states, in a matter-of-fact way, that entropy "is the uncertainty in one's knowledge of the internal configuration of the body." (1973, p. 2339) Clearly, Jaynes's information-theoretic subjectivist interpretation of statistical mechanics had won out by the 1970s – entropy has transformed from a quantity that keeps track of the reversibility or irreversibility of thermodynamic processes to a quantity that keeps track of the amount of information we have (or have lost) about said processes.

This makes Jaynes's paper the perfect candidate for evaluating the degeneration in the transition from entropy as a thermodynamic, physical, concept about physical systems to an

[23] See e.g. Brillouin (1956, Ch. 12 and beyond).

information-theoretic, subjective, concept about our ignorance or partial knowledge about physical systems.

Before Jaynes, the Gibbsian approach to statistical mechanics was by far the dominant paradigm in physics. Under the Gibbsian approach, the Gibbs entropy S_G of a physical system is defined by:

$$S_G = -k \int_{\Gamma} \rho(x,t)\ log\ \rho(x,t)\ dx \tag{1.3}$$

Here, is the $6N$-dimensional phase space of the physical system in question, x is a point in Γ, dx is the volume element of Γ, and, importantly, $\rho(x,t)$ is some probability distribution defined over Γ which may or may not change over time. The Gibbs entropy is thus a function of these probability distributions. To define $\rho(x,t)$, consider a fictitious infinite ensemble of systems having (generally differing) microstates (position and momentum) consistent with the known macrostates (e.g. temperature, volume, pressure) of the actual system. In short, the macrostate(s), together with the dynamics of course, determine the choice of $\rho(x,t)$. S_G, in turn, is supposed to match the thermodynamic entropy at the thermodynamic limit,[24] which justifies its definition and also provides a physical basis for using it. Just as the (change in) thermodynamic entropy tracked the reversibility or irreversibility of thermodynamic processes, being equal to zero for reversible processes and greater than zero for irreversible ones (for closed systems over time), so does S_G at the appropriate limit.

Setting aside the debate over the nature of these fictitious ensembles (among other conceptual issues with the Gibbsian approach) for present purposes,[25] the approach so far is physical and world-oriented. In particular, the probability distributions $\rho(x,t)$ depend on the physical state of the system and are empirically determined – for instance, a system in equilibrium with an arbitrarily large heat bath is described with the canonical ensemble distribution, an isolated system with constant energy is described with the microcanonical

[24] This is when the number of particles in the system and the volume of the system itself approach infinity, with the ratio between particle number and volume held constant.

[25] See, again, Callender (1999) and Goldstein et al. (2020) for a nice overview of the issues with the Gibbsian approach

ensemble distribution, and so on. These distributions then tell us the probability of some set of microstates obtaining given said constraints. We are here concerned simply with whether (and how likely) certain microstates of the actual physical system occur. Nothing in the Gibbsian theory *forces* us to employ notions of ignorance or knowledge so far, i.e. notions that would typically be described as 'subjective' do not need to be employed in Gibbsian statistical mechanics.[26]

In contrast, Jaynes (1957) explicitly introduces the notion of "subjective statistical mechanics" – where the usual rules of statistical mechanics can be "justified independently of any physical argument, and in particular independently of experimental verification". (1957, p. 620) For Jaynes, statistical mechanics should not be interpreted as a physical theory in itself, with its equations, choice of distributions, and rules of computation justified by physical reasoning. Rather, it should be interpreted as a system of statistical inference, concerned primarily with our partial knowledge about physical systems. This system is then underpinned by the maximum entropy principle, which prescribes maximizing entropy as a formal means of representing maximal ignorance about that which we do not know. This principle is intended by Jaynes as an a priori principle of reasoning. The physics provides only the means of enumerating the possible states of the system and their properties. (1957, p. 620) We use these as constraints on our knowledge (or lack thereof) and infer from these a set of equations via an appeal to information theory and subjectivist interpretations of probability and entropy.

Jaynes makes two related but distinct claims about his proposed account of statistical mechanics. First, Jaynes' first overarching claim is that statistical mechanics should be interpreted in a *subjectivist* fashion. This is opposed to an *objectivist* approach, which treats the probabilities produced by statistical mechanics as objective chances about events in the world (independent of what we think about those events). In his words:

[26] There is a debate about whether the use of coarse-graining in Gibbsian statistical mechanics (which is necessary to recover the second law) might be anthropocentric and hence 'subjective' – see Robertson (2020) for a discussion of why that is not necessarily the case (and hence why the Gibbsian approach still need not appeal to subjectivist notions).

> The ”subjective” school of thought regards probabilities as expressions of human ignorance; the probability of an event is merely a formal expression of our expectation that the event will or did occur, based on whatever information is available. (1957, p. 622)

For the subjectivist interpretation of statistical mechanics, the probability distributions, such as the canonical ensemble, the grand canonical ensemble, the microcanonical ensemble etc. are used to represent our partial knowledge of the system given certain constraints. The probabilities given by these ensembles are not really about the objective chances of the microstates occurring per se. Rather, these probabilities are interpreted to represent the degrees of belief we ought to have about these microstates, given suitable constraints.

What is important, however, is that the suitable constraints are not merely physical ones given by how the system is set up or behaving. This is his second major claim.

In addition to the subjectivist interpretation of statistical mechanics, Jaynes famously proposed an additional constraint to the inferential process: the maximum entropy principle (or MAXENT). (1957, p. 623) In short, the principle calls for the maximization of the entropy of the system, in addition to the other relevant physical constraints. However, the entropy of the system is interpreted in an information-theoretic way, over and above the subjectivist interpretation. Recall the Gibbs entropy:

$$S_G = -k \int_\Gamma \rho(x,t)\ log\ \rho(x,t)\ dx \tag{1.4}$$

Under the subjectivist interpretation, $\rho(x,t)$ now represents the degrees of belief we ought to have in some (set of) microstates x obtaining at time t. Over and above that, S_G is to be interpreted (with the Boltzmann constant k set to unity via a choice of units) as the (continuous version of) Shannon entropy instead, representing ‘uncertainty’ contained in $\rho(x,t)$, i.e. the ‘uncertainty’ contained within our degrees of belief about said system. The intuition is that a peaked distribution contains less uncertainty than a flat distribution, and it turns out that S_G for a peaked distribution is indeed lower than a flat distribution (see Fig. 1. for a visual aid).

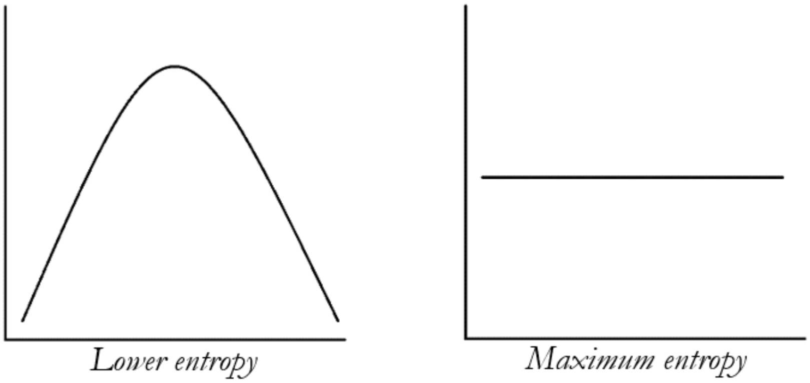

Figure 1.1. A schematic representation of a 'peaked' distribution vs. a 'flat' one.

Furthermore, just as collecting information is additive, so too is S_G additive (as a simple result of its logarithmic form).

In sum, MAXENT is treated as an a priori rationality constraint: we assume maximal ignorance about a system except what we know about it, where 'maximal ignorance' is equated to adopting a maximum information-entropy constraint on the system. This is what Jaynes meant by his account of statistical mechanics being a general account of statistical inference. (1957, p. 621, pp. 629–630) Jaynes then showed that the adoption of MAXENT can recover all the usual equations and expressions of statistical mechanics.

As with Lakatos's account in **MSRP**, much has already been said about the content of Jaynes's claims, and I will not add more to the mix.[27] I focus on the methodological *depth* of Jaynes's paper instead by applying the extended account of growth and degeneration.

I begin with the problem-situation. The founding motivations of statistical physics, found in the works of Boltzmann and Gibbs, are quite clear: to understand the molecular foundations of thermodynamics, and to interpret thermodynamics in terms of molecular mechanics.[28] As Boltzmann states in the introduction to his Lectures on Gas Theory:

[27] See e.g. Seidenfeld (1986) and the references therein.

[28] The transition from Boltzmann to Gibbs is itself an interesting chapter in the history of statistical physics, but will be left, hopefully, to future work.

> I hope to prove in the following that the mechanical analogy between the facts on which the second law of thermodynamics is based, and the statistical laws of motion of gas molecules, is also more than a mere superficial resemblance. (1896/1995, p. 5)

Gibbs, while differing in his approach to statistical mechanics, held a similar view about the goal of statistical mechanics:

> We may [...] confidently believe that nothing will more conduce to the clear apprehension of the relation of thermodynamics to rational mechanics, and to the interpretation of observed phenomena with reference to their evidence respecting the molecular constitution of bodies, than the study of the fundamental notions and principles of that department of mechanics to which thermodynamics is especially related. (1902, p. ix)

The search for an appropriate *interpretation* of statistical mechanics which connects thermodynamics to statistical mechanics was a prime focus of both Gibbs and Boltzmann, even though their methods differed significantly.

This problem-situation – the search for molecular foundations for thermodynamics – remains relevant today. For instance, Callender (2001, p. 540) notes that "kinetic theory and statistical mechanics are in part attempts to explain the success of thermodynamics in terms of the basic mechanics." More recently, Frigg and Werndl (2021) argues that this problem-situation is tied to a demand for explanation, namely why and how statistical mechanics relates to thermodynamics at all. To ignore this problem-situation is akin to quietism about this relationship, and:

> While practitioners may find it expedient to avoid the issue in this way, from a foundational point of view quietism is a deeply unsatisfactory position because it leaves the relation between [statistical mechanics] and [thermodynamics] (or indeed any macroscopic account of a system's behaviour) unexplained. (2021, p. 7)

This highlights a *need* to explain why thermodynamics is related to statistical mechanics, and that is precisely what the problem-situation is about. Indeed, Jaynes (1957, p. 620) situates his work in terms of this problem-situation as well. Suffice to say, this problem-situation is not an arbitrary one, but one that has motivated the foundations of statistical mechanics and plays a crucial role in its history and development.[29]

As is well-known, the Gibbsian approach and its associated entropy face several conceptual troubles in attaining this goal. Recall the unclear nature of the fictitious ensembles and how to interpret the probabilities provided by the approach, as well as issues such as the requirement to assume various physical hypotheses. For instance, ergodicity or metric transitivity[30] is typically introduced as a founding assumption[31] in statistical mechanics textbooks to connect measured (i.e. time) averages to expectation values over phase space. However, there are some crucial issues with this assumption: for instance, systems relevant to statistical mechanics are not always ergodic.[32] Jaynes does discuss some of these worries, such as the appropriate interpretation of statistical mechanical probabilities and the requirement for Gibbsian statistical mechanics to include seemingly arbitrary 'physical hypotheses' such as ergodicity or an a priori principle of indifference. (1957, p. 621) However, the question is whether his paper contributes in a significant way to this problem-situation. Does his proposed interpretation of entropy-as-ignorance solve the issues faced by this problem-situation? Does it provides a clearer understanding of the issues above, or does it obfuscate the issues at hand?

The stated motivations for his paper (1957, p. 261) suggest preliminary grounds for concern about superfluity. Jaynes claims that his primary motivations are (i) bringing in new

[29] Of course, it may be the case that there are other problem-situations beyond the above. Wallace has recently argued that "it may be misleading to regard statistical mechanics itself as itself wholly or primarily a conceptual underpinning for thermodynamics." (2015, 286) Instead, he opts for an approach that emphasizes an understanding statistical mechanics as used in contemporary practice, "concerned little with providing a foundational underpinning for the general principles of thermodynamics" (2015, 292), contra Frigg and Werndl above. This may well be true, and it is not the goal of this paper to argue for the 'one true problem-situation' for statistical mechanics and entropy (if one should even exist). There is no reason why degeneration must be understood only in terms of some problem-situations and not others: we could always perform the analysis of degeneration in terms of Wallace's proposed problem-situation. This raises an interesting further question of when and how problem-situations evolve and change. There could very well be an analysis of degeneration of problem-situations themselves, though I will leave that for future work. In any case, to my knowledge, Wallace's proposal remains a minority view. The goal of searching for the molecular foundations of thermodynamics remains generally accepted as historically and conceptually important. Many thanks to an anonymous reviewer for pointing this out.

[30] Roughly, ergodicity (or what Jaynes referred to as 'metric transitivity', and also known elsewhere as 'metric indecomposibility' (see Sklar (1993, 165)) is the idea that the phase space for a system is such that a phase point's trajectory is not confined only to one sub-region – another way to think about it is that the phase point's trajectory traverses the entire volume of phase space.

[31] See Frigg and Werndl (2021) for an overview of the status of this principle and related notions.

[32] See Earman and Rédei (1996).

mathematical machinery to statistical mechanics, and (ii) the notion that information theory is “felt by many people to be of great significance for statistical mechanics”, although “the exact way in which it should be applied has remained obscure.” But these motivations do not help with respect to the problem-situation above. Jaynes also does not specify any concrete problem-situation relevant to statistical mechanics, only mentioning the above issues in passing.

1.4.2 Assessing the Depth of Jaynes’s Claims: Superfluity

The distinction between subjective and objective interpretations of probabilities was presumably an attempt by Jaynes to address the issue with interpreting the probabilities prescribed by Gibbsian statistical mechanics. This is, of course, a real interpretative issue with statistical mechanics. As mentioned, we can distinguish between interpreting probabilities as worldly objective chances, or subjective degrees of belief about events (which may or may not be subject to further rational constraints); the usual debate ensues as to which interpretation is appropriate.[33] However, regardless of the result of that debate, Jaynes’s actual proposal with regards to MAXENT simply does not rely on a choice between them. As I see it, Jaynes’s proposed ‘subjectivist statistical mechanics’ is simply generalization for generalization’s sake.

The discussion of the subjective/objective distinction is *superfluous* in the context of Jaynes’s paper. To see the irrelevance and superfluity of that distinction, consider that concepts like information and uncertainty, and what is sometimes confusingly called ‘knowledge’ or ‘our knowledge’ about something, in information theory, are in fact neutral between the two interpretations of probability (contrary to folk usage of these terms). The Shannon entropy is a formal quantity that tracks the flatness of any probability distribution (be it a distribution for objective chances or degrees of belief): the more peaked it is, the more information (and less entropy) it contains. Indeed, looking at its use in communications, the relevant distributions involved are typically distributions of objective frequencies (e.g. of letters, words and so on), not degrees of belief. Unless one is understandably tricked by the occurrence of subjective-sounding

[33] For a more detailed discussion of the general positions one might take, see Sklar (1993, p. ch. 3).

words like ‘surprise’, ‘information’ (in the sense that it informs someone), ‘uncertainty’ and ‘knowledge’, the notion of information is, I claim, neutral between the objective and subjective interpretations of probability.

And why should it? Information theory is an extremely useful tool for our everyday communications, but it is ultimately a mathematical tool, without explicit metaphysical import. The distinction between objective and subjective interpretations of probability, on the contrary, is clearly a metaphysical one. As Jaynes notes: “the theories of subjective and objective probability are mathematically identical”, though they differ conceptually. (1957, p. 622) So it is for the information-theoretic (Shannon) entropy. Even though common introductions gloss it as a measure of ‘uncertainty’, this does not force a subjective interpretation of probability onto the Shannon entropy.

If so, the MAXENT proposal – which simply requires maximizing the Shannon entropy of any probability distribution, over and above other physical constraints – is likewise neutral between interpretations. We can certainly *choose* to interpret MAXENT in a subjectivist way as Jaynes did. Since we are considering only probability distributions as degrees of belief, maximizing entropy is akin to adopting the ‘flattest’ distribution of degrees of belief regarding a certain class of events given the available constraints. But we can also consider probability distributions as objective chances, in which case the MAXENT proposal becomes one in which we postulate that the probabilistic behavior of systems simply act in a way that maximizes entropy given the constraints.

Jaynes seems to see the latter as unpalatable and the former acceptable: he mentions how his subjectivist proposal avoids ‘arbitrary assumptions’ (1957, p. 630) or ‘physical hypotheses’ (1957, p. 621) several times. But why should physical hypotheses be avoided or labelled arbitrary in the field we call *physics*? Ergodicity might have its own conceptual issues and concerns over applicability, but it is surely a valid hypothesis to be considered and debated about, rather than dismissed seemingly a priori as one would in Jaynes’s approach. This is especially since ergodicity allows us to reproduce much of the physics we care about at the macroscopic level.

And, at the very least, we are making a claim about the system's actual behavior (which may obtain or otherwise) and why it fits the predictions we make about it in our theory. Compare this to the subjectivist proposal, in which the theory of statistical mechanics no longer describe the dynamics of the chances of events occurring on phase space, but merely our degrees of belief about those events occurring. As Albert famously quipped,

> Can anybody seriously think that our merely being *ignorant* of the exact micro-conditions of thermodynamic systems plays some part in *bringing it about,* in *making it the case*, that (say) *milk dissolves in coffee*? How could that be? What can all those guys have been up to? (2000, p. 64)

Is the subjectivist proposal really any less arbitrary when it comes to connecting our physical theories to the world? Jaynes does not elaborate. I do not want to adjudicate the debate here, though it suffices to say that the interpretation of probabilities is simply superfluous to the actual MAXENT proposal – the proposal itself, as a piece of mathematics, is independent of interpretation. Since both interpretations will inevitably reproduce the same mathematics (and hence the same equations), both interpretations inevitably rise and fall together.

There is little reason to think that Jaynes meant the MAXENT proposal to be much more than just a useful piece of mathematics that can help us compute and make predictions in a more tractable fashion, for it seems that his discussion of MAXENT entirely brackets off the issue of interpretation. If that is the case, however, the question of interpreting probabilities in statistical mechanics does not even arise. The actual goal of the paper is not about interpretation or the metaphysics of statistical mechanics. In turn, the question of interpreting the thermodynamic entropy, defined over these probabilities *about the system*, does not arise. Rather, MAXENT is a proposal concerning convenient prediction and computation. Jaynes writes:

> Although the principle of maximum-entropy inference appears capable of handling most of the prediction problems of statistical mechanics, it is to be noted that prediction is only one of the functions of statistical mechanics. Equally important is the problem of interpretation; given certain observed behavior of a system, what conclusions can we draw as to the microscopic causes of that behavior? To treat this problem and others like it, a different theory, which we may call objective statistical mechanics, is needed. (1957, p. 627)

The MAXENT proposal is here just a convenient proposal for arriving at the computations required for prediction problems. Hence Jaynes claimed that adopting the 'subjective point of view' and MAXENT for predictions serves a "great practical convenience". But if we were to press Jaynes on the interpretation and metaphysics of statistical mechanics, we would still need 'objective statistical mechanics':

> In the problem of interpretation, one will, of course, consider the probabilities of different states in the objective sense; i.e., the probability of state n is the fraction of the time that the system spends in state n. (1957, p. 627)

Jaynes's take on the interpretation of probabilities about the actual physical system remains an objectivist one – and one seemingly adopting some version of the ergodic hypothesis he claimed to have eschewed![34] This goes to show that the probabilities prescribed by statistical mechanics about actual systems – and their interpretations – are not even in question here in Jaynes's paper, since his proposal is supposed to be one concerning 'subjective statistical mechanics', rather than 'objective statistical mechanics'. If we are only interested in prediction and computation, all we need is the *mathematical* MAXENT proposal, and a formal proof that it does in fact recover the equations we want. The ability for the MAXENT proposal to shorten and speed up the derivations of certain equations (as Jaynes shows in the paper) is, by and large, not in question here. Yet there is also no need to provide an interpretation for MAXENT and the notion of entropy involved in that case. There is no more need to interpret the mathematical shortcuts that one takes, any more than one needs to justify and interpret the algorithms behind WolframAlpha when one takes a shortcut with their integrals. All this renders Jaynes's insistence on packaging the MAXENT proposal with a choice of interpretation for both probabilities and entropy confusing.

Furthermore, since we are not tackling the *actual* issue of how to interpret the proba-

[34] As Frigg and Werndl (2021, p. 8) puts it, ergodicity holds "if for all measurable functions the infinite time average is equal to the ensemble average for almost all initial conditions". Roughly, an ergodic principle states that the time-averages of some parameter is equal (in the infinite time limit) to the expectation value of that parameter (and hence the probability of that parameter having some quantity is equal to that amount of time the system spent in the region of phase space with that quantity, as Jaynes says here).

bilities assigned to the states of the actual system, the subjectivist MAXENT package is not even *relevant* to the original problem-situation. In other words, the insistence on providing an information-theoretic interpretation of entropy – replacing the previous thermodynamic and physical interpretation via the second law of thermodynamics and notions of reversibility/irreversibility of actual processes – is simply unjustified because the MAXENT proposal has no real need for interpretation.

No other genuine argument for the information-theoretic interpretation can be found in his paper. He starts off with a proviso:

> The mere fact that the same mathematical expression $\sum p_i \ log \ p_i$ occurs both in statistical mechanics and in information theory does not in itself establish any connection between these fields. This can be done only by finding new viewpoints from which thermodynamic entropy and information-theory entropy appear as the same *concept*. In this paper we suggest a reinterpretation of statistical mechanics which accomplishes this, so that information theory can be applied to the problem of justification of statistical mechanics. (1957, p. 621)

As I have shown, information theory is not ultimately applied to the *justification of statistical mechanics*. That project requires interpreting statistical mechanics and the metaphysics within it (e.g. about whether swarms of particles can actually recover the macroscopic description). Despite Jaynes's claim that he is proposing a reinterpretation of statistical mechanics, he does not succeed in doing so – that is all left in the 'objective statistical mechanics' side of things, which he has chosen to downplay. Jaynes's proposal is the adoption of new mathematical tools for computing predictions in statistical mechanics, which does not force any interpretation at all. In any case, such interpretations have no real import for the actual issue of the problem-situation, that of interpreting the probabilities attached to events themselves. It is important to note that Jaynes does not specify an alternative problem-situation either. Instead, he simply asserts:

> Since $\sum p_i \ log \ p_i$ is just the expression for entropy as found in statistical mechanics, it will be called the entropy of the probability distribution p_i; henceforth we will consider the terms "entropy" and "uncertainty" as synonymous. (1957, p. 622)

But he has not yet shown that the thermodynamic entropy, i.e. "entropy", and the information-theoretic entropy, i.e. "uncertainty", are the same as a matter of interpretation, because the paper is not at all concerned with interpretation and 'objective statistical mechanics', only prediction and 'subjective statistical mechanics'.

In sum, Jaynes's discussion of interpretative issues is *superfluous*. Jaynes has added unnecessary terminology from information theory and confused these new concepts with old questions without actually addressing any of the old questions from the original problem-situation. He has generalized for generalization's sake. The insistence on interpreting entropy as information-theoretic ignorance in a subjectivist sense, defined over distributions interpreted as degrees of belief, is likewise superfluous. As Denbigh correctly notes, "Jaynes' remark [on interpreting entropy in a subjectivist manner], though undoubtedly illuminating in a certain sense, is quite superfluous to the actual scientific discussion". (1990, p. 111)

1.4.3 Assessing the Depth of Jaynes's Claims: Authoritarianism

Jaynes also displays authoritarianism when insisting on treating statistical mechanics as a general means of prediction, apparently viewed through subjectivist lens.

Authoritarians introduce new concepts into a line of inquiry without justification, while ignoring the problem-situation and heuristics which led us to those concepts. This is an important issue in Jaynes's paper, since the problem-situation at hand is barely specified. No details about the issues facing 'objective statistical mechanics' or the Gibbsian approach are presented. Instead, Jaynes presents the information-theoretic interpretation, the subjectivist interpretation, and the maximum entropy principle, as though they must be taken altogether.

Jaynes proclaims that, in

> freeing [statistical mechanics] from its apparent dependence on physical hypotheses of the above type, we make it possible to see statistical mechanics in a much more general light. (1957, p. 621)

Throughout the paper, Jaynes insists that the subjectivist approach is necessary for approaching the prediction issue. However, two questions arise. First, why the downplaying of ‘physical hypotheses’ used by ‘objective statistical mechanics’ and why do we need to ‘free’ statistical mechanics from them? Second, why the focus on prediction and information theory, and the downplaying of the importance of interpretation? Both questions are unanswered.

To the first question, Jaynes demands that a satisfactory theory connecting microscopic to macroscopic phenomena should, among other things, “involve no additional arbitrary assumptions”. (1957, pp. 620–621) He notes a worry that this condition might be too severe since, rightfully, “we expect that a physical theory will involve certain unproved assumptions, whose consequences are deduced and compared with experiment.” (1957, 621) However, his response to this worry is unsatisfactory. After listing some additional assumptions historically used in statistical mechanics to ensure empirical adequacy, he notes that

> with the development of quantum mechanics the originally arbitrary assumptions are now seen as necessary consequences of the laws of physics. This suggests the possibility that we have now reached a state where statistical mechanics is no longer dependent on physical hypotheses, but may become merely an example of statistical inference. (1957, 621)

However, the fact that *some* arbitrary assumptions eventually come to be explained by quantum mechanics does not entail that statistical mechanics is (or should be) free of all physical hypotheses. The possibility that this could be possible is not a good argument for thinking that this is in fact the case (which would be necessary for him to argue so strongly against the use of physical hypotheses). Furthermore, even granting that this were true, the inference from this to statistical mechanics becoming “merely an example of statistical inference” is an unexplained leap as well.

In short, Jaynes believes that these physical hypotheses are undesirable – for instance, he spends some time claiming that metric transitivity is not needed if we adopt the MAXENT principle. (1957, p. 624) However, he never provides an adequate reason for why we should not adopt any physical hypotheses about the systems we are studying.

He focuses instead on how MAXENT can help us do away with these hypotheses. But it is not clear that it does – MAXENT merely shifts our attention away from whether those hypotheses hold. Jaynes notes that adopting MAXENT is in fact akin to adopting ergodicity – except about our own degrees of belief about the system's behavior, rather than about the actual system's behavior:

> Even if we had a clear proof that a system is not metrically transitive, we would still have no rational basis for excluding any region of phase space that is allowed by the information available to us. In its effect on our ultimate predictions, this fact [i.e. MAXENT] is equivalent to an ergodic hypothesis, quite independently of whether physical systems are in fact ergodic. (1957, p. 624)

Is the system *really* ergodic? And is ergodicity needed to derive the equations concerning those systems' behavior? Jaynes's proposal has two options: one is to say nothing at all – an unsatisfactory answer. Another option is to reply: the MAXENT proposal says that you should have degrees of belief matching the situation where the system is ergodic (as the quote above suggests). But that means I ought to believe that the system is ergodic after all, i.e. believing the *physical hypothesis of ergodicity*. Yet that was the original issue in our problem-situation: we want to know whether ergodicity is necessary for the statistical mechanical system to behave in accordance with our observations. Either MAXENT is irrelevant to our problem-situation, or it adds nothing new. Old questions remain.

The original problem-situation has been neglected. Yet, we are made to believe that these questions are to be ignored in favour of the new proposal – MAXENT, information theory, subjectivism – without justification for why that should be so. This is a case of authoritarianism.

Turning to the second question: as discussed above, the founding fathers of statistical mechanics were concerned first and foremost with the interpretative issues – how do we connect the particles or systems of statistical mechanics to the bulk macroscopic behavior we find in thermodynamics? Of course, that is not to say that prediction has no role to play in statistical mechanics. However, it is strange to ignore a core tenet of statistical mechanics, which seems like what Jaynes has done here. Reading the paper, one gets the impression that prediction holds supreme place in statistical mechanics. Interpretation seems to be an after-thought. But prediction goes hand in hand with interpretation – to predict the behavior of the system we must understand what the system is, and that is a matter of interpretation. As I have argued in §1.4.2, Jaynes's paper is completely divorced from interpretative issues. In this respect his paper is authoritarian: it ignores the problem-situation of statistical mechanics, such as the importance of interpretative issues.

The introduction of information theory, and the shift in focus on statistical mechanics as a general tool of statistical inference, is likewise authoritarian. Jaynes offers no reason for adopting information theory – we are told that the Gibbs entropy can be interpreted as the Shannon entropy, and that "the development of information theory has been felt by many people to be of great significance for statistical mechanics". (1957, p. 621) Likewise, we are not told why statistical mechanics should be a general tool of statistical inference, freed from physics, where "the usual rules are thus justified independently of any physical argument, and in particular independently of experimental verification." (1957, p. 620) These are all core tenets of the MAXENT proposal, but they remain unjustified.

In conclusion, Jaynes's paper falls afoul of both superfluity and authoritarianism. With respect to methodological depth, then, it was a degenerative piece of work. Since the key transition of entropy from a concept concerned with thermodynamics and actual physical systems to a concept concerned with ignorance and our knowledge of said systems occurred here, this shift is a degenerative one as well.

Jaynes' paper changed the trajectory of the entropy concept. For instance, by appearing as though the paper presented an interpretative package, despite the actual proposal not needing one, the paper introduced confusion to the actual interpretative issues. The interpretative package of information-theoretic entropy and subjectivism became adopted as an answer for the interpretative issues associated with 'objective statistical mechanics' instead. Recall the Bekenstein quote: a mere twenty years later the information-theoretic interpretation has escaped from 'subjective statistical mechanics' into 'objective statistical mechanics', with *thermodynamic* entropy (defined over probability distributions of the microstates of the actual *systems*) being interpreted as ignorance or uncertainty. Three years later, Hawking would simply assert: "an intimate connection between holes (black or white) and thermodynamics [...] arises because information is lost down the hole." (1976, p. 197) Degeneration has occurred.

1.4.4 Content-Oriented Degeneration

It is worthwhile to conclude by briefly considering the content-degeneration of the MAXENT proposal under the generalized Lakatosian account I have developed in III. It seems to me that MAXENT is both theoretically and empirically degenerative on this account.

Recall the definitions: being theoretically progressive refers to a succeeding theory predicting more novel facts compared to its predecessor. Being empirically progressive refers to the excess empirical content of this succeeding theory actually leading to the discovery of new facts, thereby corroborating the new theory's novel predictions.

As Jaynes notes, nothing new is added in terms of theoretical progress, because 'subjective statistical mechanics' will *recover exactly the same predictions* as 'objective statistical mechanics':

> Conventional arguments, which exploit all that is known about the laws of physics, in particular the constants of the motion, lead to exactly the same predictions that one obtains directly from maximizing the entropy. (1957, p. 624)

And that

> the subjective theory leads to exactly the same predictions that one has attempted to justify in the objective sense. (1957, p. 625)

This shows that no new predictions are provided by this proposal. The 'new' proposals attached in Jaynes's papers are typically just new ways of doing the same calculations. For instance: Jaynes's treatment of Siegert's 'pressure ensemble' is also merely a reworking of Siegert's own derivations, published just a year prior. (1956) In short, the MAXENT proposal is theoretically degenerative. Furthermore, since there are no new predictions, there are no new predictions to corroborate. The proposal is empirically degenerative. This, of course, also adds to that sense of superfluity one gets when analyzing Jaynes's proposal in detail.

Overall, then, Jaynes's paper is degenerative tout court in terms of its place in statistical mechanics.[35] Given that it had such a huge influence on the current understanding of entropy as ignorance and uncertainty, especially in the field of contemporary black hole thermodynamics (Bekenstein and Hawking are typically known as the founding fathers of black hole thermodynamics), this current understanding must be re-assessed. Some have already begun this work. For instance, Wüthrich have recently argued that

> the original argument by Bekenstein with its detour through information theory does not succeed in establishing the physical salience of the otherwise merely formal analogy between thermodynamic entropy and the black hole area, and so cannot offer the basis for accepting black hole thermodynamics as "the only really solid piece of information". (2018, pp. 219–220)

Importantly, Wüthrich diagnoses the problem with Bekenstein's arguments to be the failure to recognize that "Fundamental physics is about the objective structure of our world, not about our beliefs or our information", and that "information, one might argue, is an inadmissible concept in fundamental physics." (2018, p. 217) Given my analysis here, we can see why that is the case. The introduction of information theory by Jaynes to statistical mechanics was already superfluous to begin with. Others like Prunkl & Timpson (ms), recognizing the flaws with information-theoretic arguments for black hole thermodynamics, are already attempting to provide a defense of black hole thermodynamics sans information theory. A possibility of doing so further suggests – in agreement with my diagnosis here – that information-theoretic concepts may just have been superfluous to the discussion, and, to quote Wuthrich again, a "detour".

1.5 Conclusion

I have provided and motivated an extension to Lakatos's account of growth and degeneration from **MSRP** by appealing to **P&R**. This extension, in terms of superfluity and

[35] There is certainly room to argue for this paper's contributions in the context of other fields, given Jayne's and MAXENT's ubiquity outside of statistical mechanics. However, this is not Jayne's original goal.

authoritarianism, enables a new dimension through which we may evaluate a piece of mathematical or scientific work, independent of the analysis in terms of theoretical and empirical progress or degeneration found in **MSRP**.

As proof of concept, I have evaluated Jaynes's proposal, a key transition point in the historical trajectory of the concept of entropy. I hope to have shown that my account does provide a novel means of assessing the degeneration or progress of this transition, by critically analyzing the aspects of his paper which exhibited superfluity and authoritarianism.

Some might object that my criticisms of Jaynes' proposal could have been made independent of the account of degeneration I have sketched here.[36] I agree that one could have arrived at these criticisms independent of my account, for there are likely many ways to arrive at the same conclusion I reached. However, that does not discount the fact that my account of degeneration does arrive at these criticisms, guided by the twin heuristics of superfluity and authoritarianism. I hope to have shown in this paper that this account provides us with a grip on the nature of these criticisms (as methodological ones) and motivates them in a conceptually clear fashion. This should give us a good reason to consider and adopt this account of degeneration regardless of whether there might be other ways to arrive at these criticisms. In any case, we should rejoice – not despair – when there are multiple ways of evaluating a problem, for this means we have more tools in our conceptual toolbox for analysis.

In my view, developing more tools for understanding how scientific and mathematical concepts degenerate or grow has natural affinities with an increasingly popular understanding of philosophy as conceptual engineering.[37] As Chalmers [2020, p. 4) writes, conceptual engineering is "the project of designing, evaluating, and implementing concepts", where we consider not only what a concept is, but also what it should be. Developing new tools for identifying points of degeneration in a concept's historical trajectory helps us evaluate a concept and consider alternative ways of designing and developing said concept.

[36] Many thanks to an anonymous reviewer who raised this point.

[37] See e.g. Haslanger (2000) or Chalmers (2020).

This paper thus leaves behind a variety of fruitful directions, ripe for the picking by the hopeful conceptual engineer. For those who, like me, are puzzled by the concept of entropy: if we want to re-engineer and design a newer, better, conceptually clearer notion of entropy, we would do well to engage with – and dispel – other similarly degenerative transition points. For other philosophers, too, I believe the tools developed here can be used to assess concepts elsewhere: in science, mathematics, perhaps even philosophy itself. There remains much to be done.

Acknowledgements

"Degeneration and Entropy", is, in full, a reprint of the material as it appears in the special issue of *Kriterion – Journal of Philosophy* 36 (2), 123-155, 2022, on *Lakatos's Undone Work: The Practical Turn and the Division of Philosophy of Mathematics and Philosophy of Science* (eds. S. Nagler, H. Pilin, and D. Sarikaya). The dissertation author was the sole author of this paper.

Bibliography

Albert, David Z. (2000). *Time and Chance*. Cambridge, Massachusetts: Harvard University Press.

Bekenstein, Jacob D. (1973). “Black Holes and Entropy”. In: *Phys. Rev. D* 7.8, pp. 2333–2346. DOI: 10.1103/PhysRevD.7.2333. URL: https://link.aps.org/doi/10.1103/PhysRevD.7.2333.

Boltzmann, Ludwig (1896–1995). *Lectures on Gas Theory*. New York: Dover.

Brillouin, Louis Marcel (1956). *Science and Information Theory*. New York: Academic Press.

Bub, Jeffrey (2005). “Quantum mechanics is about quantum information”. In: *Foundations of Physics* 35.4, pp. 541–560.

Callender, Craig (1999). “Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy”. In: *The Journal of Philosophy* 96.7, pp. 348–373.

— (2001). “Taking Thermodynamics Too Seriously”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32.4. The Conceptual Foundations of Statistical Physics, pp. 539–553. ISSN: 1355-2198. DOI: https://doi.org/10.1016/S1355-2198(01)00025-9. URL: https://www.sciencedirect.com/science/article/pii/S1355219801000259.

Chalmers, David (2020). “What is conceptual engineering and what should it be?” In: *Inquiry*. DOI: 10.1080/0020174X.2020.1817141.

Chua, Eugene Y. S. (2021). “Does Von Neumann Entropy Correspond to Thermodynamic Entropy?” In: *Philosophy of Science* 88.1. DOI: 10.1086/710072.

Corfield, David (1997). “Assaying Lakatos’s philosophy of mathematics”. In: *Studies in History and Philosophy of Science* 28, pp. 99–121.

Denbigh, Kenneth (1990). "How Subjective is Entropy?" In: *Maxwell's demon: Entropy, Information, Computing*. Ed. by Harvey Leff and Andrew Rex. Princeton, New Jersey.

Dougherty, John and Craig Callender (2016). *Black Hole Thermodynamics: More Than an Analogy?* Available at: http://philsci-archive.pitt.edu/13195/. (last accessed 13th March 2022).

Earman, John and Miklós Rédei (1996). "Why Ergodic Theory Does Not Explain the Success of Equilibrium Statistical Mechanics". In: *The British Journal for the Philosophy of Science* 47, pp. 63–78.

Frigg, Roman and Charlotte Werndl (2021). "Can Somebody Please Say What Gibbsian Statistical Mechanics Says?" In: *The British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axy057.

Friston, Karl and Klaas E. Stephan (2007). "Free-energy and the brain". In: *Synthese* 159.3, pp. 417–458.

Gibbs, Josiah W. (1902). *Elementary principles of statistical mechanics: Developed with special reference to the rational foundation of thermodynamics*. New Haven: Yale University Press.

Goldstein, Sheldon et al. (2020). "Gibbs and Boltzmann Entropy in Classical and Quantum Mechanics". In: *Statistical Mechanics and Scientific Explanation*. Ed. by Valia Allori. Singapore: World Scientific, pp. 519–581.

Hallett, Michael (1979). "Towards a Theory of Mathematical Research Programmes II". In: *The British Journal for the Philosophy of Science* 30.2, pp. 135–159.

— (n.d.). "Towards a Theory of Mathematical Research Programmes I". In: *The British Journal for the Philosophy of Science* 1.1979a (), pp. 1–25.

Haslanger, Sally (2000). "Gender and Race: (What) Are They? (What) Do We Want Them To Be?" In: *Noûs* 34.1, pp. 31–55.

Hawking, Stephen (1976). "Black holes and thermodynamics". In: *Physical Review D* 13.2, pp. 191–197.

Jaynes, Edward T. (1957). “Information Theory and Statistical Mechanics”. In: *Physical Review* 106.4, pp. 620–630.

Kampis, George, Ladislav Kvasz, and Michael Stöltzner (2002). *Appraising Lakatos: Mathematics, methodology, and the man.* Dordrecht: Kluwer.

Kiss, Olga (2006). “Heuristic, Methodology or Logic of Discovery? Lakatos on Patterns of Thinking”. In: *Perspectives on Science* 14.3, pp. 302–317.

Lakatos, Imre (1976–2015). *Proofs and refutations: The logic of mathematical discovery.* Ed. by J. Worrall and E. Zahar. Cambridge: Cambridge University Press.

— (1978). *The methodology of scientific research programmes.* Ed. by J. Worrall and G. Currie. Vol. 1. New York: Cambridge University Press.

Leng, Mary (2002). “Phenomenology and mathematical practice”. In: *Philosophia Mathematica* 10, pp. 3–25.

Lewis, M.B. and A.J.F. Siegert (1956). “Extension of the Condensation Theory of Yang and Lee to the Pressure Ensemble”. In: *Physical Review* 101.4, pp. 1227–1233.

McIrvine, Edward C. and Myron Tribus (Sept. 1971). “Energy and Information”. In: *Scientific American.* URL: https://www.scientificamerican.com/article/energy-and-information/.

Natsuume, Makoto (2015). *AdS/CFT Duality User Guide in Lecture Notes in Physics 903.* Tokyo: Springer.

Prunkl, Carina and Christopher Timpson (n.d.). *Black Hole Entropy is Thermodynamic Entropy.* DOI: https://doi.org/10.48550/arXiv.1903.06276.

Roach, Ty N.F. (2020). “Use and Abuse of Entropy in Biology: A Case for Caliber”. In: *Entropy* 22.12, p. 1335. DOI: 10.3390/e22121335.

Robertson, Katie (2020). “Asymmetry, Abstraction, and Autonomy: Justifying Coarse-Graining in Statistical Mechanics”. In: *The British Journal for the Philosophy of Science* 71.2, pp. 547–579. DOI: 10.1093/bjps/axy020.

Seidenfeld, Teddy (1986). “Entropy and Uncertainty”. In: *Philosophy of Science* 53.4, pp. 467–491.

Sklar, Lawrence (1993). *Physics and Chance: Philosophical issues in the foundations of statistical mechanics*. Cambridge: Cambridge University Press.

Stöltzner, Michael (2002). "What Lakatos Could Teach the Mathematical Physicist". In: *Appraising Lakatos: Mathematics, methodology, and the man*. Ed. by George Kampis, Ladislav Kvasz, and Michael Stöltzner. Dordrecht: Kluwer.

Swinburne, James (1904). *Entropy, or, Thermodynamics from an Engineer's Standpoint: And the Reversibility of Thermodynamics*. Westminster: Constable.

Uffink, Jos (2001). "Bluff Your Way in the Second Law of Thermodynamics". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32.3, pp. 305–394. DOI: 10.1016/s1355-2198(01)00016-8.

Wallace, David (2015). "The quantitative content of statistical mechanics"". In: *Studies in History and Philosophy of Modern Physics* 52, pp. 285–293.

— (2020). "The Necessity of Gibbsian Statistical Mechanics". In: *Statistical Mechanics and Scientific Explanation*. Chap. Chapter 15, pp. 583–616. DOI: 10.1142/9789811211720_0015. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789811211720_0015. URL: https://www.worldscientific.com/doi/abs/10.1142/9789811211720_0015.

Werndl, Charlotte (2009). "Justifying definitions in mathematics: going beyond Lakatos". In: *Philosophia Mathematica* 17.3, pp. 313–340.

Wüthrich, Christian (2018). *Are Black Holes About Information?" In: Why Trust A Theory?: Epistemology of Fundamental Physics*. Ed. by Radin Dardashti, Richard Dawid, and Karim Thebault. New York, NY: Cambridge University Press, pp. 202–223.

Chapter 2

Does Von Neumann Entropy Correspond to Thermodynamic Entropy?

2.1 Introduction

According to conventional wisdom in physics, von Neumann entropy corresponds to phenomenological thermodynamic entropy. The origin of this claim is von Neumann's (1955) argument that his proposed entropy corresponds to the thermodynamic entropy, which appears to be the only explicit argument for the equivalence of the two entropies. However, Hemmo and Shenker (**H**&**S**) (2006) – and earlier, Shenker (1999) – have argued that this correspondence fails, contrary to von Neumann. If so, this leaves conventional wisdom without explicit justification.

Correspondence can be understood, at the very least, as a numerical consistency check: in this context, this means that the von Neumann entropy has to be included in calculating thermodynamic entropy to ensure consistent accounting in contexts where both thermodynamic and von Neumann entropy are physically relevant. Successful correspondence provides strong evidence of equivalence. While it does not guarantee equivalence, it seems to be at least a necessary condition for equivalence. If thermodynamic entropy and von Neumann entropy correspond, then we have reason to think that von Neumann entropy is rightfully thermodynamic in nature, since proper accounting of thermodynamic entropy would demand von Neumann entropy. By contrast, a failure of correspondence seems to entail that the von Neumann entropy is *not* thermodynamical in nature, since it is irrelevant to thermodynamic calculations in contexts where both entropies are physically significant (e.g. when a system has both quantum degrees of freedom *and* is sufficiently large to warrant thermodynamical considerations).

Although Henderson (2003), in my view, has successfully criticized Shenker's earlier argument, little has been done in the philosophical literature to evaluate **H**&**S**'s more recent arguments.[1] This lacuna is striking because, as I mentioned, von Neumann's argument appears to be the only explicit argument for correspondence for the two entropies.

My goal in this paper is to fill this lacuna by providing a novel set of criticisms to **H**&**S**. Here's the plan: I introduce key terms (§2.2) and then present von Neumann's thought-experiment which aims to establish the correspondence between thermodynamic entropy and von Neumann entropy; along the way, a novel counterpart to the usual argument for correspondence is discussed (§2.3). I then present and criticize **H**&**S**'s arguments for the single-particle case in the context of thermodynamics (§2.4.1) and in the context of statistical mechanics (§2.4.2), the N-particles case (§2.4.3), and the infinite-particles case (§2.4.4). I conclude that their argument fails in all cases – in turn, we have good reasons to reject their claim that the von Neumann entropy fails to correspond to thermodynamic entropy, and hence the claim that von Neumann entropy is not thermodynamic in nature.

2.2 Key Terms

Let me first define the notions of thermodynamic entropy and von Neumann entropy. Following **H&S**, I define the *change in thermodynamic entropy* S_{TD} between two thermodynamic

[1] It is only slightly better in the physics literature: Deville and Deville (2013) appears to be the only paper to critique **H**&**S**. On the philosophical side, one (very recent) exception is Prunkl (ms), though she restricts discussion to the single-particle case and appears to conflate information entropy with thermodynamic entropy. See §2.4.1/§2.4.2 for why this is not obviously right.

states in an *isothermal quasi-static* process,[2] as:

$$\Delta S_{TD} = \frac{1}{T}\int P\,dV \tag{2.1}$$

We will restrict our discussion to ideal gases in equilibrium (i.e. systems where pressure P, volume V, and temperature T remain constant).

Next, the *von Neumann entropy* S_{VN}, for any pure or mixed quantum system, is defined as:

$$S_{\text{VN}} = -kTr(\rho\ log\ \rho) \tag{2.2}$$

where k is the Boltzmann constant and Tr(.) is the trace function. Generally, the density matrix ρ is such that:

$$\rho = \sum_{n=1}^{i} p_i \left|\psi_i\right\rangle \left\langle\psi_i\right| \tag{2.3}$$

where $\psi_1, \psi_2, ...\ \psi_n$ correspond to the number of pure states in a statistical mixture represented by ρ, with $p_1, p_2, ...\ p_n$ being their associated classical probabilities (which must sum to unity). In the case where there is only one pure state possible for a system (e.g. when we are absolutely certain about its quantum state), then n = 1, with probability 1, so the appropriate density matrix is $\rho = \left|\psi\right\rangle\left\langle\psi\right|$. For such a system in a pure state (i.e. represented by a single state vector in Hilbert space), $S_{\text{VN}} = 0$. For mixed states (i.e. states which cannot be represented by a single state vector in Hilbert space, hence *mixture* of pure states or a *mixed state*), Tr(ρ log ρ) < 1 and $S_{\text{VN}} > 0$ in general. A mixed state is often said to represent our *ignorance* about a system – this will suffice as a first approximation (more on how to interpret this ignorance in §2.4.2).

Prima facie, S_{VN} and S_{TD} appear to share nothing in common, apart from the word 'entropy'. However, von Neumann claims that there are important correlations between the two, which suggests a correspondence between S_{TD} and S_{VN}.

[2] There is no change in temperature in an isothermal quasi-static process, which is why T is taken to be constant. As a matter of historical note, von Neumann uses an isothermal set-up in his argument, with a box containing a quantum ideal gas coupled to a (much larger) heat sink ensuring constant temperature over time (von Neumann 1955, 361/371).

2.3 Von Neumann's Thought-Experiment

For parity, I adopt **H**&**S**'s presentation of von Neumann's thought-experiment,[3] which aims to show that changes in thermodynamic entropy can only be made consistent with the laws of thermodynamics if we considered the von Neumann entropy as contributing to the calculation of the thermodynamic entropy. Figure 2.1 depicts the stages of the thought-experiment.

We begin, in stage one, with a box with a partition in the middle. On one side of the partition there is a gas at volume V, constant temperature T, and constant pressure P. Each gas particle starts off having the pure state spin-up along the x-direction $\left|\psi_x^{\uparrow}\right\rangle$, which is equivalent to a superposition of spin-up and spin-down pure states along the z-direction, labelled $\left|\psi_z^{\uparrow}\right\rangle$ and $\left|\psi_z^{\downarrow}\right\rangle$ respectively. According to standard quantum mechanics, the state of each particle is thus $\frac{1}{\sqrt{2}}(\left|\psi_z^{\uparrow}\right\rangle + \left|\psi_z^{\downarrow}\right\rangle)$.

In this context, particles with quantum behavior may be taken to be *ideal gases*, i.e. sets of particles each of which do not interact with other particles and take up infinitesimal space. Following von Neumann's assumptions (von Neumann 1955, 361),[4] each gas particle is understood as a quantum particle with a spin degree of freedom contained inside a large impenetrable box, and each gas particle is put inside an even larger container isolated from the environment (i.e. the box we began with). This ensures that each spin degree of freedom is incapable of interacting with other particles. These boxes' sizes also ensure that the positions of these boxes (and hence of the particles) can be approximately classical. Since the container is much larger than each gas particle, this ensures that the gas particles take up negligible space relative to the massive container. Accepting these assumptions, we may then take these quantum particles to behave like an ideal gas.[5] Following **H**&**S**, we further assume that the

[3] It is not clear to me that von Neumann's original 1932/1955 argument is exactly the same as the argument **H**&**S** reproduces. However, for the sake of argument, I will refer to **H**&**S**'s version as von Neumann's argument in this paper.

[4] These assumptions are borrowed from Einstein (1914). For more, see Peres (Peres 2002, 271).

[5] I shall follow everyone in this debate in assuming that the above set-up is physically possible.

position degrees of freedom of the gas particles have no interaction with the spin degrees of freedom at this point, and "due to the large mass of the boxes, the position degrees of freedom of the gas may be taken to be classical and represented by a quantum mechanical mixture". (Hemmo and Shenker 2006, 155)

Moving on, stage two involves a spin measurement along the z-axis on all the particles in the container, with a result being an equally weighted statistical mixture of particles with either $\left|\psi_z^\uparrow\right\rangle$ or $\left|\psi_z^\downarrow\right\rangle$ states. As a result, the spin state of each particle is then represented instead by a density matrix ρ_{spin}, such that:

$$\rho_{spin} = \frac{1}{2}(\left|\psi_z^\uparrow\right\rangle \left\langle\psi_z^\uparrow\right| + \left|\psi_z^\downarrow\right\rangle \left\langle\psi_z^\downarrow\right|) \tag{2.4}$$

More precisely, there should be terms for the measurement device too, when truly considering the entire system. ρ_{spin} describes only the subsystem (i.e. the quantum ideal gas) *sans* measurement device, i.e. a state with the measurement device traced out – this is in line with von Neumann's focus on the entropy changes due to changes in the subsystem (von Neumann 1955, 358–379). I follow Henderson (2003) and **H&S** in talking about the system's state as though I have already traced the measurement device out whenever measurement is involved.

Stages three and four are where the particles are (reversibly) separated according to their spin states by a semi-permeable wall into two sides of the box, each with volume V.[6] As a result of this separation, we in effect double the mixture's volume. The gas expands to fill up volume V on each side.

Stage five involves an isothermal and quasi-static compression of the mixture so that we return to a total volume V (effectively halving the volume on each side of the box), while pressure on both sides becomes equal. Importantly, due to this compression, S_{TD} decreases due to the decrease in volume.

[6] This semi-permeable wall can be assumed to be a black box which reversibly separates particles to different sides based on their different orthogonal/disjoint states; see (von Neumann 1955, 367–370) for discussion. I follow everyone in the debate in accepting this assumption.

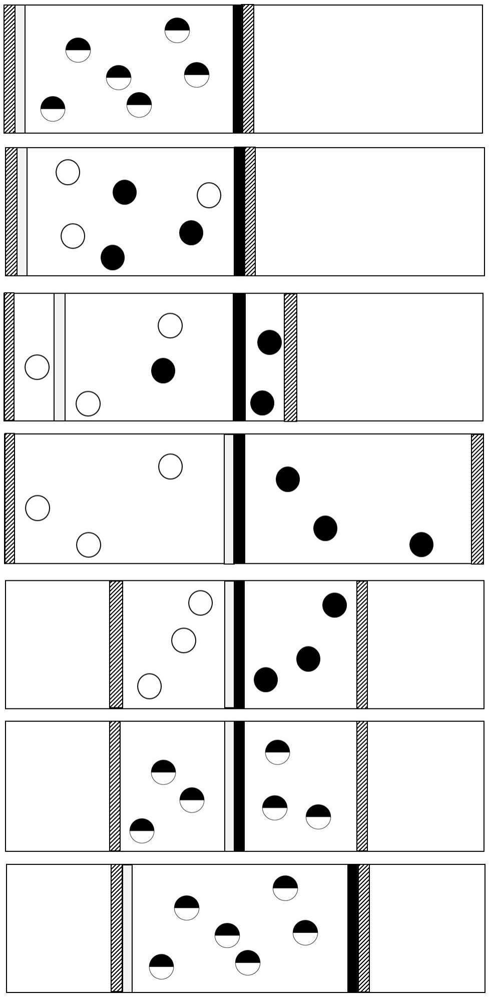

Figure 2.1. From top to bottom, stage one to stage seven, as described by **H**&**S**.

Stage six brings all the particles into the pure spin state $\left|\psi_x^{\uparrow}\right\rangle$ quasi-statically and without work done, while stage seven removes the semi-permeable wall, such that the system returns to its original state.

Now consider how S_{VN} and S_{TD} change across the various stages. Stage seven ends with the body of gas having the same thermodynamic state (same V, same P, and constant T) as stage one. Furthermore, all the thermodynamic transformations performed were reversible, and removing the wall alone does no additional work. Thus, the system at stage one must have the same thermodynamic entropy as stage seven, i.e. $\Delta S_{\text{TD}} = 0$, since S_{TD} depends only on the initial and final state of the system. $\Delta S_{\mathrm{VN}} = 0$ from stage one to seven too, since the system is in the same state in both the first and seventh stages.

Since stage six does not involve thermodynamic transformations, there is no change in S_{TD}. Likewise, the transformation of ρ_{spin} to $\left|\psi_x^\uparrow\right\rangle$ here does not change S_{VN} as the transformation can be performed unitarily. This is possible as a result of our separation of the gases to different sides of the box according to their spin-eigenstates - given this, we can perform unitary operations on each side of the box (or perform the more general measurement procedure recommended by (von Neumann 1955, 365-367)), to transform them into the same state as stage one. Both unitary transformations and von Neumann's procedure do not increase S_{VN}, and so there is no change in S_{VN} at stage six as a result.

There are no changes in S_{TD} or S_{VN} in stages three and four. While there is an increase in the gas's volume, as noted above, from V to $2V$, and hence an accompanying increase in S_{TD} by $n.R.log\ 2$,[7] there is also a compensating change in the thermodynamic entropy of mixing[8,9] by $-n.R.log\ 2$ which exactly compensates this increase in S_{TD}.[10] Since the particles are in orthogonal spin states at this stage, there are no quantum effects (e.g. 'collapse' effects) from simply filtering the gases with the semi-permeable walls, and hence S_{VN} does not change either.[11]

[7] Here, n refers to the number of moles of gas in the system, and R is the gas constant.

[8] Henderson (2003) explains the mixing entropy, describing the mixing of different gases, crisply: "After separation, each separated gas occupies the original volume V alone. To return to the mixture, each gas is compressed to a volume $c_i V$ (where c is the concentration of the i^{th} gas). The compression requires work $W = -n.k.T \sum_i\ c_i\ log\ c_i$ to be invested, and the entropy of the gas is reduced by $\Delta S = -n.k. \sum_i\ c_i\ log\ c_i$. An increase in entropy of the same amount must then be associated with the mixing step of removing the partitions. This is the 'mixing entropy'." (Henderson 2003, 292) Separation simply results in a decrease in entropy of the same amount.

[9] Tim Maudlin raised the following objection to the applicability of the entropy of mixing in this context when a version of this paper was presented at a summer school. Mixing should have a thermodynamic effect only when differences between the gases are already assumed to be thermodynamically relevant: for example, mixing differently colored gases should not have a thermodynamic effect unless the difference in color is thermodynamically relevant. It is, however, not clear whether the difference in spin is a thermodynamically relevant one, and might amount to begging the question. This is a good point, but one that I am setting aside for now, since everyone in the debate accepts the assumption that separating the gases here decreases the entropy of mixing. As we shall see later, a more fundamental issue arises with using the entropy of mixing in the 'single particle' case.

[10] see (**H&S** 2006, 157, fn. 4).

[11] This is argued for in (von Neumann 1955, 370–376).

However, importantly, there is a decrease in S_{TD} in stage five, of $-n.R.log\ 2$ due to the isothermal compression and decrease in volume. Yet, nowhere else is there any further change in S_{TD}. We have to account for why the overall change in S_{TD} from the first to the seventh stages is 0.

As von Neumann argues, only one possibility remains. While S_{TD} remains constant in stage two, notice that there was an increase in S_{VN}, of $-N.k.-log\ 2 = \frac{N.R}{N_A}.log\ 2 = n.R.log\ 2$,[12] as a result of the spin measurement. This is equivalent to the change of S_{TD} in stage five. The state of each particle changes from a pure state $\frac{1}{\sqrt{2}}(|\psi_z^{\uparrow}\rangle + |\psi_z^{\downarrow}\rangle)$ to a mixed state represented by ρ_{spin}, and hence S_{VN} for the gas increases on the whole. In order to ensure that entropic changes are consistent, von Neumann thinks that we should accept S_{VN}'s contribution to S_{TD} in this context, where both quantum effects and thermodynamical considerations are at play. Without accepting S_{VN} in our entropic accounting, we end up with a violation of thermodynamics since we have a reversible thermodynamic cycle with non-zero change in S_{TD}, *contra* the Second Law. In other words, we should accept that S_{VN} corresponds to S_{TD}.

Furthermore, the correspondence of S_{VN} and S_{TD} in this context can be defended from another perspective, apart from considerations about consistency from the thermodynamic perspective: consistent accounting from the perspective of *quantum mechanics* also demands correspondence. This is simply a change in perspective with regards to the thought-experiment, but, to my knowledge, this argument has not been explicitly made in the literature, thus underselling the case for correspondence in von Neumann's thought experiment.

Instead of arguing for correspondence by considering thermodynamic consistency, i.e. ensuring that ΔS_{TD} = 0 throughout the cycle, we can also consider consistency from the quantum mechanical perspective. We started and ended with the same spin state, and so it should be the case that $\Delta S_{\text{VN}} = 0$ throughout the cycle. Yet, there is an inconsistency: if we only consider the increase of S_{VN} in stage two as a result of measurement, we should end in

[12] N is the total number of particles: since each particle is assumed to be non-interacting and independent from others under the ideal gas assumption, their entropies are additive. N_A is Avogadro's number.

stage seven with an *increase* in S_{VN}, *not* $\Delta S_{\text{VN}} = 0$. As described, there is nowhere else in the thought-experiment where S_{VN} changes. However, there *is* a decrease in S_{TD} in stage five due to the *thermodynamic process* of isothermal compression, exactly balancing out the increase in S_{VN}. Hence, we can ensure consistency, i.e. that $\Delta S_{\text{VN}} = 0$, only by taking S_{VN} to correspond to S_{TD}. In other words, just as the thermodynamic accounting of S_{TD} is consistent only if we consider S_{VN}, the quantum entropic accounting of S_{VN} is *also* consistent only if we consider S_{TD}. Consistency from a quantum mechanical perspective also demands correspondence between S_{VN} and S_{TD}.

Though the debate has largely focused only on how the thought-experiment demonstrates one direction of correspondence, of S_{VN} *to* S_{TD} as a result of thermodynamical considerations, the correspondence demonstrated by this thought-experiment in fact goes *both ways.* Of course, since von Neumann was focused on demonstrating the *thermodynamic* nature of S_{VN} (specifically the irreversibility of measurement), rather than the *quantum* nature of S_{TD}, it was natural that he chose to approach it the way he did.

2.4 Hemmo and Shenker's Arguments

H&S disagree with von Neumann's argument, and criticize it by considering three cases: the single-particle case, the finite but large N particles case, and the infinite particles case.

2.4.1 Single Particle Case - Thermodynamics

H&S first consider von Neumann's argument in the single particle case (see Figure 2.2). They claim that the argument does not go through here, since S_{TD} actually remains constant, contrary to our thought-experiment's description. In other words, using thermodynamical considerations, they find that S_{VN} should not be included in our accounting for S_{TD}.

Here's their argument. Consider the stages where there are entropic changes. In stage two when the spin measurement was performed, S_{VN} increases as before, since it tracks the change of the particle's spin state from pure to mixed.

Contrariwise, S_{TD} does not change in stage five (isothermal quasi-static compression) *nor anywhere else* (this will be important later). After stage two, the single particle is in either the $\left|\psi_z^{\uparrow}\right\rangle$ state or the $\left|\psi_z^{\downarrow}\right\rangle$ state. After stages three and four, with the expansion and separation via semi-permeable wall, there is a particle only in *one* side of the box, and not the other. We make an S_{TD}-conserving location measurement[13] to figure out which side of the box is empty and which side the particle is at, so as to compress the box against the empty side. The compression is then performed as per before. However, this compression does *not* decrease S_{TD}:[14] to restore the volume of the 'gas' to V no work needs to be done, since we are compressing against vacuum. Since there is a change in S_{VN} in this cycle, but no change in S_{TD}, the apparent answer, in order to do our entropic accounting, is to *ignore*, not incorporate, S_{VN} into S_{TD}. Hence S_{VN} does not correspond to S_{TD}.

Their analysis is problematic. Though their ultimate point in this analysis – that S_{TD} fails to corresponds to S_{VN} – still holds, it does not hold in the way they claim. In fact, the way it fails suggests to us that we should *disregard* the single particle case.

For the single particle case, they claim that "... [S_{TD}] is null throughout the experiment." (**H**&**S** 2006, 162) This then allows them to claim that thermodynamic accounting for S_{TD} is consistent *only if* we did not consider S_{VN}. This then supports their claim that S_{VN} does not correspond to S_{TD} since adding S_{VN} into the thermodynamic accounting actually renders the otherwise consistent calculations inconsistent.

They are right to say that the stage five compression (after location measurement) has no thermodynamic effect because we are compressing against vacuum: no work needs to be

[13] Prunkl (ms) claims that the location measurement leads to a violation of the Second Law. If true, this makes **H**&**S**'s argument even more problematic. Here, for the sake of argument, I assume that the location measurement is unproblematic.

[14] As an anonymous reviewer rightfully notes, the location measurement is important for ensuring $\Delta S_{\text{TD}} = 0$ here. Without the location measurement, we might end up compressing in the wrong direction against the side with the gas, rather than the empty vacuum - this will have thermodynamic effects since we are doing work on the gas. However, the **H**&**S** set-up emphasizes the location measurement, and I will play along for the sake of argument.

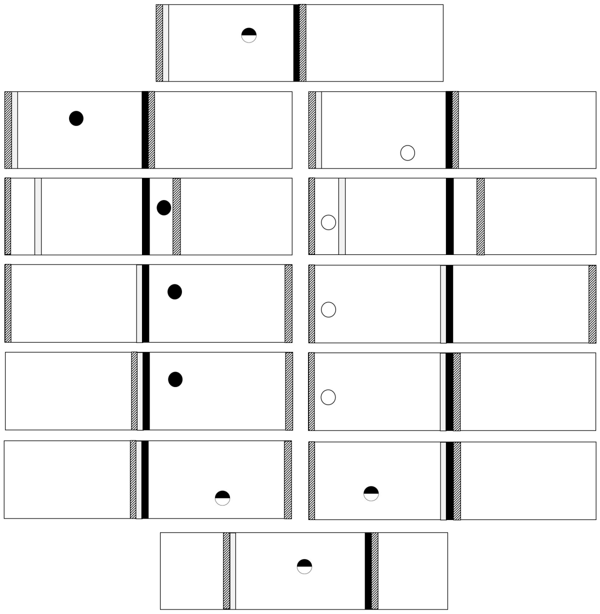

Figure 2.2. From top to bottom, stage one to stage seven for the single particle case as described by **H**&**S**.

done, and so $\Delta S_{\text{TD}} = 0$ for stage five. *However*, I claim that $\Delta S_{\text{TD}} \neq 0$ for the single particle case *overall*, because $\Delta S_{\text{TD}} \neq 0$ in stages three and four in this context.

As far as I can tell, **H**&**S** did not analyze stage three and four, i.e. the isothermal expansion and separation, in terms of the single particle case at all. Rather, they seem to have assumed that $\Delta S_{\text{TD}} = 0$ in these stages as with the original case of the macroscopic gas.[15] However, this assumes that there is both a change in entropy of $n.R.log\ 2$ due to isothermal expansion *and* a change in the entropy of mixing of $-n.R.log\ 2$ due to separation, as they say so themselves for the original case: "The increase of thermodynamic entropy due to the volume increase $\Delta S = \frac{1}{T} \int P\ dV$ is exactly compensated by the decrease of thermodynamic mixing entropy $\Delta S = \sum w_k\ ln\ w_k$ (where w_k is *the relative frequency of molecules of type k*) due to the separation." (**H**&**S** 2006, 157, fn. 4, emphasis mine)

In the single particle case, it makes sense that isothermal expansion should still increase S_{TD}, since the single particle 'gas' is expanding against a piston and doing work. However,

[15] Prunkl (ms) appears to do the same.

it *does not make physical sense* to speak of the entropy of mixing here at all, since *there is no separation of gases* in the single particle case. The entropy of mixing is explicitly defined for systems where *different* gases are separated from/mixed with one another via semi-permeable walls, but a single particle cannot be separated from/mixed with itself. The quote above makes this conceptual point explicit: by **H**&**S**'s own lights, the relative frequency of a single particle is simply unity (and null for particles of other types), so the entropy of mixing is $1 \; ln \; 1 = 0$. *There is no thermodynamic entropy of mixing in the single particle case.*

Discounting the entropy of mixing, however, we find that $\Delta S_{\text{TD}} = n.R.log \; 2 \neq 0$ for stages three and four, and hence for the entire process, contrary to **H**&**S**'s claim. Interestingly, correspondence *does* fail to obtain between S_{TD} and S_{VN}, since $\Delta S_{\text{TD}} + \Delta S_{\text{VN}} = 2n.R.log \; 2 \neq 0$, despite the process being reversible *ex hypothesi*: incorporating S_{VN} into thermodynamic accounting violates the Second Law.

However, on this new analysis, we gain some clarity as to why the single particle case is problematic. While it is true that incorporating S_{VN} into the thermodynamic accounting violates the Second Law, S_{TD} accounting *by itself* also violates the Second Law (contrary to **H**&**S**). Even without considering S_{VN}, $\Delta S_{\text{TD}} \neq 0$ despite the process being reversible. Thermodynamic accounting is inconsistent here *no matter what we do*, which suggests that the reversible process they described for the single particle case is thermodynamically unsound: if so, any argument **H**&**S** make in this context may be disregarded.

The upshot: I agree with **H**&**S** that correspondence fails for the single particle case, but not *why* it fails. It is *not* because the process they described is already thermodynamically consistent without taking S_{VN} into account. Rather, it is because the process is already thermodynamically *inconsistent* anyway.

In recent work, John Norton argued that thermodynamically reversible processes for single-particle systems are impossible in principle, which might explain *why* the process described by **H**&**S** is thermodynamically unsound: it was not justified to assume the process was reversible for a single particle system. For Norton, a reversible process is "loosely speaking, one whose driving forces are so delicately balanced around equilibrium that only a very slight disturbance to them can lead the process to reverse direction. Because such processes are arbitrarily close to a perfect balance of driving forces, they proceed arbitrarily slowly while their states remain arbitrarily close to equilibrium states." (Norton 2017, 135) Norton notes that these thermodynamic equilibrium states are balanced not because there are no fluctuations, but because these fluctuations are negligible for macroscopic systems. However, fluctuations relative to single-particle systems are large, and generally prevent these systems from being in equilibrium states at any point of the process, rendering reversible processes impossible in the single particle case. (Norton 2017, 135) If reversible processes are impossible for single particle systems in general, then it should come as no surprise that the particular single particle reversible process used by **H&S** is likewise thermodynamically unsound, as my analysis above suggests. If so, **H&S**'s claim that correspondence fails in this process is simply besides the point, since this process is not thermodynamic at all.

Since any reversible process cannot be realized for single particle systems in general, the issue seems not to be with any particular process *per se*, but with the single particle case *simpliciter*. To my knowledge, no one prior to **H&S** discussed von Neumann's experiment in terms of a single particle; von Neumann (1955), Peres (1990, 2002), Shenker (1999) and Henderson (2003) all explicitly or implicitly assume a large (or infinite) number of particles. This is for good reason. As **H&S** acknowledge, and as we have seen: "The case of a single particle is known to be problematic as far as arguments in thermodynamics are concerned". (**H&S** 2006, 158) Matter in phenomenological thermodynamics is assumed to be continuous.[16] A 'gas' composed of one particle can be many things, but it is surely not continuous in any commonly accepted sense. In other words, it is just not clear whether the domain of thermodynamics should apply to the single-particle case at all.

As Myrvold (2011) notes, Maxwell also made a similar claim with regards to phenomeno-

[16] See Compagner (1989) for a discussion of the so-called 'continuum limit' as a counterpart to the thermodynamic limit in phenomenological thermodynamics.

logical thermodynamics *in general*; it does not and should not hold in the single particle case. On his view, the laws of phenomenological thermodynamics, notably the Second Law, must be continually violated on small scales:

> If we restrict our attention to any one molecule of the system, we shall find its motion changing at every encounter in a most irregular manner.
>
> If we go on to consider a finite number of molecules, even if the system to which they belong contains an infinite number, the average properties of this group, though subject to smaller variations than those of a single molecule, are still every now and then deviating very considerably from the theoretical mean of the whole system, because the molecules which form the group do not submit their procedure as individuals to the laws which prescribe the behaviour of the average or mean molecule.
>
> Hence the second law of thermodynamics is continually being violated, and that to a considerable extent, in any sufficiently small group of molecules belonging to a real body. As the number of molecules in the group is increased, the deviations from the mean of the whole become smaller and less frequent [...] (Maxwell 1878, 280)

The Second Law, and hence phenomenological thermodynamics, should not be expected to hold true universally in small scale cases, and especially *not* in the single-particle case. Von Neumann and everyone else in the debate should have recognized this point. Why, then, should it matter that the thought-experiment succeeds or fails in this case? Phenomenological thermodynamics does not apply to single-particle cases. There is thus no profit in trying to establish correspondence between S_{VN} and S_{TD} in this case. Indeed, if we took seriously Maxwell's claim that the Second Law fails at small scales, a failure of thermodynamic entropic accounting might even be *expected*; it does not rule out the possible thermodynamic nature of S_{VN} even though the sum of S_{VN} and S_{TD} might be inconsistent with the Second Law. In short, it is not clear *why* the single-particle case is relevant to the discussion at hand.

H&**S**'s reasoning is untenable, because they fail to respect the context of phenomenological thermodynamics by bringing it into a context where it is not expected to hold. Instead, it seems more appropriate that the single-particle case is precisely beyond the purview of classical thermodynamics, requiring an analogue that only corresponds to classical thermodynamics at the appropriate scales and limits. We may then take S_{VN} to be the analogue of S_{TD} in this case, only approximating S_{TD} as the system in question approaches the context suitable for traditional thermodynamic analysis. If so, we may see von Neumann as merely demonstrating that S_{VN} corresponds, not at *all* domains but *in the domain where thermodynamics is taken to hold*, to S_{TD}.

2.4.2 Single Particle Case Redux - Statistical Mechanics and Information

Given the foregoing discussion, **H**&**S** might insist that S_{VN} fails to correspond to S_{TD} *even when* take into account a more relevant domain for single particles – statistical mechanics.

After directly arguing that S_{VN} does not correspond to S_{TD} (**H**&**S** 2006, 162–165), they further argue that S_{VN} does not correspond to information entropy (more on this below) in the single-particle case. *Prima facie*, this should seem *irrelevant* to von Neumann's argument, which was to establish the correspondence of the *thermodynamic* S_{TD} and *quantum* S_{VN}: why should *information* entropy's failure to correspond with S_{VN} be a worry at all?

Here's one plausible worry, on a charitable reading. *If* information entropy corresponds to S_{TD}, *and* **H**&**S** shows that S_{VN} fails to correspond to information entropy, then we might conclude, indirectly, that S_{VN} does not correspond to S_{TD} after all.[17] This argument assumes that information entropy *does* correspond to S_{TD}, an assumption **H**&**S** seem to hold as well: this is in line with the so-called 'subjectivist' view of statistical mechanics.[18] Furthermore, my above argument against the misapplication of phenomenological thermodynamics does not seem to apply here, since this argument is being made in the context of statistical mechanics and its particle picture, with no commitment to phenomenological thermodynamics.

However, **H**&**S** do not do much to motivate the linkage between information entropy and S_{TD}; indeed, in their words, "a linkage between the Shannon information and thermody-

[17] Caveat: I am not committed to the information entropy's relationship to thermodynamics. One may, like Earman and Norton (1998, 1999), be skeptical that information entropy is related to S_{TD} at all, in which case **H**&**S**'s argument here is simply irrelevant.

[18] Notably, see Jaynes (1957).

namic entropy has not been established" (**H**&**S** 2006, 164). Without this link, the failure of correspondence between the information entropy and S_{VN} appears, at best, irrelevant to the correspondence between S_{TD} and S_{VN}. Nevertheless, I will take a charitable view here and assume that there *is* a correspondence between information entropy and S_{TD}, for the sake of assessing their argument. Here's a plausible (if arguable) sketch: if one were a subjectivist like Jaynes (1957), one might take the Gibbs entropy in statistical mechanics to be a special case of the information entropy. After all, both have the form:

$$-\sum_{i} p_i \, ln \, p_i \tag{2.5}$$

with *i* being the number of possible states with associated probabilities of occurring p_i, with the Gibbs entropy being multiplied by an additional Boltzmann's constant k.[19] We know that statistical mechanics corresponds to phenomenological thermodynamics at the thermodynamic limit so we can think of the Gibbs entropy, and hence information entropy, as corresponding to S_{TD}. I take this to be in line with what **H**&**S** have in mind: "to the extent that the Shannon information underwrites the thermodynamic entropy, it does so via statistical mechanics" (2006, 165). *Assuming* that the above picture is plausible, a failure of correspondence between S_{VN} and the information entropy provides evidence against the correspondence between S_{VN} and S_{TD}.

Their argument comes into two parts. Ignoring S_{TD} for the time being (which does not change throughout the cycle for the single-particle case – see §2.4.1), they claim that we can consider the stage five location measurement to be a decrease in *information entropy* of $ln \; 2$, as a result of learning information about which one of two parts of the box the particle is in. On first glance, this seems to resolve the arithmetic inconsistency in entropic accounting: $ln \; 2$ is exactly the increase in S_{VN} as a result of the spin state changing from a pure state $\left|\psi_x^{\uparrow}\right\rangle$ to the mixed state ρ_{spin} in stage two. In other words, for both the information and von Neumann

[19] Using the so-called Planck units, where $k = 1$, Gibbs entropy and information entropy are then formally equivalent.

entropy's accounting to be correct (i.e. net change of zero across the cycle), we must consider S_{VN} as corresponding to information entropy. Now, since information entropy also corresponds, *ex hypothesi*, to S_{TD}, we have an indirect argument for the correspondence of S_{VN} to S_{TD}.

However, **H**&**S** claim that this argument fails for *collapse* interpretations, i.e. interpretations of quantum mechanics on which a superposed quantum state ontologically collapses into a pure state upon measurement (either precisely or approximately).[20] They allow that, on *no-collapse* interpretations, e.g. Bohmian or many-worlds interpretations, the location measurement in stage five does not decrease S_{VN}, since the state of the system never changes in light of measurements, and so the above argument goes through.

Let us see what they could mean by this claim by following the state of the particle through the cycle. At stage two, everyone agrees that the state of the system is ρ_{spin} following the z-spin measurement; S_{VN} increases by $ln\ 2$. At this point, the particle's position degrees of freedom remain independent from its spin degrees of freedom, as per our ideal gas assumption, though we might assume the particle starts out on the left half of the box, with the mixture of position states $\rho_{pos}(L)$ with 'L' representing the left side. (Consider Figure 3.1 but with only one particle). Following the semipermeable wall's filtering at stages three and four, the location of the particle becomes classically correlated with the spin. Let's say that the semipermeable wall sends $\left|\psi_z^{\uparrow}\right\rangle$ particles to the left, represented by $\rho_{pos}(L)$, and $\left|\psi_z^{\downarrow}\right\rangle$ particles to the right, represented by $\rho_{pos}(R)$. As such, the (mixed) state of the particle is now:

$$\rho_{particle} = \frac{1}{2}\Bigg(\left|\psi_z^{\uparrow}\right\rangle \left\langle\psi_z^{\uparrow}\right| \otimes \rho_{pos}(L) + \left|\psi_z^{\downarrow}\right\rangle \left\langle\psi_z^{\downarrow}\right| \otimes \rho_{pos}(R)\Bigg) \tag{2.6}$$

For no-collapse interpretations, **H**&**S** agree that the state of the particle stays the same as above after the location measurement in stage five. We perform the compression in stage five and remove the partition at the end of stage six, thereby removing the classical correlations

[20] On GRW-type approaches, though, collapse occurs with or without measurement, but measurement increases the likelihood of collapse, roughly speaking.

between position and spin. No further change in either information entropy or S_{VN} occurs, and hence the correspondence goes through (**H**&**S** 2006, 164) - the spin state remains mixed until unitarily transformed into a pure state and completing the cycle.

For collapse interpretations, they claim that the location measurement *decreases* S_{VN} by ln 2, because, on collapse interpretations, the state of the particle upon the measurement, depending on which side the particle is found, becomes:

$$\rho_{particle} = \begin{cases} \left|\psi_z^{\uparrow}\right\rangle \left\langle\psi_z^{\uparrow}\right| \otimes \rho_{pos}(L) \\ \left|\psi_z^{\downarrow}\right\rangle \left\langle\psi_z^{\downarrow}\right| \otimes \rho_{pos}(R) \end{cases} \tag{2.7}$$

The spin state of the system here effectively goes from being a mixed state to a pure state as a result of this measurement: S_{VN} *decreases* by ln 2. Summing up the entropy changes, there was a decrease of ln 2 in information entropy, and a net change of zero for S_{VN} as a result of the increase in stage two and the decrease in stage five. Overall, then, the change is *not* zero but $-ln$ 2; our accounting has gone awry, and there is a failure of correspondence between S_{VN} and information entropy. If this is right, S_{VN} does not correspond to S_{TD}.

However, I think that **H**&**S** are wrong to claim that S_{VN} decreases following the location measurement for collapse interpretations. As Prunkl (Prunkl ms, 11–12) notes, there is an inconsistency here. Everyone, *including* **H**&**S**, agrees that the spin state of the particle is mixed – *not* pure – after stage two's spin measurement, *even on* collapse interpretations (**H**&**S** 2006, 160). In that case, why does the particle's spin become *pure* after the location measurement?

I think this results from a confusion over the nature of mixed states. In particular, they seem to have adopted what Hughes (Hughes 1992, §5.4, §5.8) call the "ignorance interpretation" of mixed states, confusing what I call *classical* and *quantum* ignorance. They seem to be assuming that mixed states simply represents *classical* ignorance, i.e. the lack of knowledge about a *particular system*: a system represented by a mixed state *really is* in a pure state, but we know not which. This is why the location measurement is supposed to reveal to us the pure state of this system (by revealing which side it is on and hence the correlated spin state) and hence 'wash away' our classical ignorance of the real state of the system - post-measurement, we know exactly which pure state *this* system is in, unlike pre-measurement; hence S_{VN} decreases.

However, as Hughes (Hughes 1992, 144/150) argues, this interpretation of mixed states – as representing classical ignorance about which pure state a *particular* system is in – cannot be the right interpretation of all mixed states. To begin, a mixed state can be decomposed in non-unique ways in general. Here's a simple example: a mixed state representing a mixture of $\left|\psi_z^\uparrow\right\rangle$ and $\left|\psi_z^\downarrow\right\rangle$ can *also* represent a mixture of $\left|\psi_x^\uparrow\right\rangle$ and $\left|\psi_x^\downarrow\right\rangle$ and so on. If we insist that a mixed state represent our classical ignorance about the real state of a particular system, then we end up having to say that a system's state is really *both* either $\left|\psi_z^\uparrow\right\rangle$ or $\left|\psi_z^\downarrow\right\rangle$, *and* either $\left|\psi_x^\uparrow\right\rangle$ or $\left|\psi_x^\downarrow\right\rangle$. Of course, this is impossible given quantum mechanics. The defender of the classical ignorance interpretation might insist that we simply pick one pair of possible pure states but not both at once. In general, however, there's no way to do that non-arbitrarily given some density matrix. Furthermore, this problem only worsens when we consider that there are usually more than just two ways to decompose a density matrix - a principled choice based on the mixed state alone is not feasible. The mixed state cannot be a representation of classical ignorance.

Instead, to paraphrase Hughes (Hughes 1992, 144–145), mixed states should be (minimally) interpreted as such: if we prepared in *the same way* an ensemble of systems, each described with the same mixed state, i.e. a mixture of pure states with certain weights, then the relative frequency of any given measurement outcome from the ensemble is exactly what we would get if the ensemble comprised of various 'sub-ensembles' each in one of the pure states in the mixture, with the relative frequency of each sub-ensemble in the ensemble given by the corresponding weights.

In other words, the sort of *quantum* ignorance relevant in the right interpretation of mixed states is not whether we are ignorant about the *real* state of *this* particular system, but whether we are ignorant about the *measured* states of an *ensemble* of *identically prepared* systems like this one. If this is right, quantum ignorance cannot be 'washed away' upon measurement of a single system unlike the sort of ignorance **H**&**S** were implicitly assuming, and it seems like this quantum ignorance is precisely what remains after the location measurement.

This was roughly Henderson's (2003) criticism against Shenker (1999), which is why it is puzzling that **H**&**S** (2006) commit the same mistake:

> This preparation produces the pure states [$\left|\psi_z^{\downarrow}\right\rangle$] and [$\left|\psi_z^{\uparrow}\right\rangle$] with equal probabilities. In a particular trial, the observer may take note of the measurement result, and he therefore discovers that he has say a [$\left|\psi_z^{\uparrow}\right\rangle$]. If he applies a projective measurement in the [$\{\left|\psi_z^{\uparrow}\right\rangle, \left|\psi_z^{\downarrow}\right\rangle\}$] basis, he could predict that he will measure [$\left|\psi_z^{\uparrow}\right\rangle$]. However, this does not mean that, if someone handed him another state prepared in the same way, he could again predict that the outcome of his measurement would be [$\left|\psi_z^{\uparrow}\right\rangle$]. In this sense the observer does not know the state of the system which is being prepared, and it is because of this ignorance that the state is mixed. Looking at the measurement result does not remove the fact that there is a probability distribution over the possible outcomes. (Henderson 2003, 294)

This applies to the location measurement in stage five too: measuring the location of the particle in *this* case does not change the state of the particle from a mixed one to a pure one even on collapse interpretations. Firstly, it seems quite irrelevant whether we adopt a collapse or no-collapse interpretation, because the collapse mechanism applies to superposed pure states, not statistical mixtures. If anything, collapse had already happened in the stage two measurement procedure, yet everyone including **H**&**S** (**H**&**S** 2006, 160) accepts that the system is in a *mixed* state after stage two even for collapse interpretations. More importantly, there remains a probability distribution over the states of the particle as a result of stage two's spin measurement, even *after* the location measurement. Given an ensemble of particles prepared from stages one to five in the same way, we are *still* not be able to predict with certainty whether an ensemble of particles would all be measured on the left or right sides of the box (and hence all spin-up or spin-down) as a result of the mixed state resulting from stage two, only that half of the ensembles will be on the left and the other half will be on the right. Quantum ignorance remains – the system remains in a mixed state even after the location measurement, as:

$$\rho_{particle} = \frac{1}{2}\Bigg(\left|\psi_z^{\uparrow}\right\rangle \left\langle\psi_z^{\uparrow}\right| \otimes \rho_{pos}(L) + \left|\psi_z^{\downarrow}\right\rangle \left\langle\psi_z^{\downarrow}\right| \otimes \rho_{pos}(R)\Bigg) \tag{2.8}$$

This is exactly the state of the system in no-collapse interpretations, i.e. quantum ignorance does *not* discern between collapse and no-collapse interpretations. What *has* gone away is the classical ignorance that **H**&**S** (mistakenly) assumed was relevant for mixed states, ignorance about *this particular system*'s state. By measuring the system's location, we come to learn of the correlations between location measurement and the particle's spin. This ignorance does not change the mixed state to a pure state: instead, this loss of classical ignorance – *gain in information* – is represented as a *decrease* in information entropy just as before, and this information is what we *use* to perform the compression in stage five.

As a result, there is no additional decrease in S_{VN} in stage five for collapse interpretations; the entropy accounting lines up after all, as with no-collapse interpretations: the decrease in information entropy *does* correspond to the increase in S_{VN}, and so information entropy does correspond to S_{VN} after all. **H**&**S**'s argument does not establish the failure of correspondence between S_{VN} and S_{TD} via the failure of S_{VN} and information entropy to correspond.

To sum up, their arguments in the single-particle case are either ill-motivated and irrelevant to von Neumann and our discussion of correspondence when considered in terms of phenomenological thermodynamics, or outright fails when considered in the more relevant domain of (informational approaches to) statistical mechanics. Either way, their argument does not support the failure of correspondence between S_{VN} and S_{TD}.[21]

[21] Let me briefly note that their argument in the two particles case fails for similar reasons. On the one hand, from the perspective of phenomenological thermodynamics, their argument is irrelevant: following Maxwell and others, two particles do not a thermodynamic system make. On the other hand, in the domain of statistical mechanics, the analysis in terms of information entropy is irrelevant from non-informational views of statistical mechanics. From an informational perspective, however, their argument rests again on the supposed difference between collapse and no-collapse interpretations of mixed states. Since this difference is non-existent, their argument likewise fails apart in that case.

2.4.3 Finitely Many Particles

H&**S**'s argument in the case of finitely many particles rests on the assumption of *equidistribution*, i.e. that the particles will be equally distributed across the left and right sides of the box after separation by the semi-permeable wall.

Assuming equidistribution, the increase in S_{VN} given the spin measurement in stage two is $Nln2$ (**H**&**S** 2006, 169). Furthermore, the decrease in thermodynamic entropy in the fourth stage is $Nln2$ as well. The entropic accounting therefore seems to work out.

However, **H**&**S** press further on the 'rough' nature of equidistribution when N is large but finite: they claim that the change in S_{VN} will only only be $Nln2$ when N is infinite, since equidistribution only truly holds when $N \to \infty$. In other cases, S_{VN} will strictly only *approximate* S_{TD}, and hence S_{VN} and S_{TD} combined will never be *exactly* zero; hence, "Von Neumann's argument goes through as an approximation" (**H**&**S** 2006, 169). However, they claim that this state of affairs suggest, instead, that von Neumann's argument strictly *fails*: "[...] since Von Neumann's argument is meant to establish a conceptual identity between the quantum mechanical entropy and thermodynamic entropy, we think that such an implication is mistaken [...] no matter how large N may be, as long as it is finite, the net change of entropy throughout the experiment will not be exactly zero." (**H**&**S** 2006, 169)
As I have already discussed in §2.4.1, it is not clear to me that von Neumann's goal really was to establish *strict identity* (what they call "conceptual identity"), i.e. correspondence between S_{VN} and S_{TD} in *all* domains. Rather, it seems to be the establishing of correspondence *only* in domains where S_{TD} is taken to hold. If so, their argument here simply misses the point.

Furthermore, as is well-known, the particle analogue of thermodynamics, statistical mechanics, become equivalent to phenomenological thermodynamics only when $N = \infty$, viz. when N arrives at the thermodynamic limit. As such, to complain that S_{VN} does not match up to S_{TD} outside of this domain is to demand the unreasonable, since it is not clear that even statistical mechanics, the bona fide particle analogue of thermodynamics, can satisfy this demand. Since

S_{VN} approximates S_{TD} the same way statistical mechanical entropies approximate S_{TD} (and becomes equivalent at $N = \infty$), and physicists generally accept that statistical mechanics corresponds to thermodynamics nevertheless, why should this problem of approximation be particularly problematic for S_{VN}? I think **H**&**S** take too seriously the notion of conceptual identity involved in von Neumann's thought-experiment to be *strict* equality, though I suspect a better way to understand von Neumann's strategy is to understand S_{VN} as an approximation to S_{TD} that is more fundamental than S_{TD} in small N cases, but becomes part of the S_{TD} calculus in domains where S_{TD} applies.

To have a case against S_{VN} as a quantum analogue of S_{TD} in the case of finitely many particles, **H**&**S** must explain what exactly the problem is with approximations in *this* case, if it has worked out so well for the case of statistical mechanics and thermodynamics. If not, they might just be "taking thermodynamics too seriously'.[22]

One might say something stronger: unless they can justify why we cannot use approximations at all in science, they do not have a case at all. As they note themselves, S_{TD} is itself only *on average approximately* $-Nln2$ (**H**&**S** 2006, 169), only being equal to $-Nln2$ when $N = \infty$. So, in fact, the approximate quantity of S_{VN}, $\sim Nln2$, exactly matches the approximate quantity of S_{TD}, $\sim -Nln2$, in the case of finitely many particles. Unless there is something wrong with approximations in physics *in general*, this, then, is in fact a case of S_{VN} corresponding to S_{TD}, contrary to their argument.

2.4.4 Infinitely Many Particles

H&**S** consider von Neumann's argument in the infinite particles case in two different ways: one as $N \to \infty$ and one as $N = \infty$. As they rightly point out, the two cases are very different for calculations of physical quantities.

Consider stage two and stage five in this context. **H**&**S** emphasize that a spin measurement is "a physical operation which takes place in time" (**H**&**S** 2006, 170), which constrains

[22] See Callender (2001).

what is physically possible.

For the case where $N \to \infty$, stage two is to be understood as a succession of physical measurements where "we measure individual quantities of each of the particles separately and only then count the relative frequencies" (**H**&**S** 2006, 170), before coming up with a density matrix describing this state. In this case, as with the case described in §2.4.3, S_{VN} approaches $Nln2$ as $N \to \infty$. Their complaint here consist of two premises: one, that, as with §2.4.3, S_{VN} never reaches $Nln2$ unless $N = \infty$. Two, that since measurements are physical measurements, we can never perform an infinite series of these measurements, and so we can never measure infinite particles. *A fortiori* the measurable S_{VN} can never arrive at $Nln2$, and so the entropic accounting is again supposed to be inconsistent if we consider both S_{VN} and S_{TD}.

However, it is clear that their argument is moot given a clear understanding of the sort of thermodynamics we are interested in (see §2.4.3). While it is true that S_{VN} will never reach $Nln2$, recall that S_{TD} (or, more likely, one of its statistical mechanical analogues, given the domain of finitely many particles merely *approaching* ∞ rather than $N = \infty$) will likewise never reach $Nln2$. In other words, it does not matter that we can never perform an infinite series of these measurements, and hence never come to know of S_{VN} at the thermodynamic limit, since we can likewise never have a thermodynamic entropy equivalent to $Nln2$ unless we are at the thermodynamic limit. The two entropies, then, in fact *correspond* in this case.

What of the second case? Here, **H**&**S** concede that "arithmetically Von Neumann's argument goes through at the infinite limit" (**H**&**S** 2006, 172), which makes sense because, as I have insisted so far, von Neumann's strategy was never to demonstrate the *strict identity* of S_{VN} and S_{TD}, i.e. the correspondence of S_{VN} and S_{TD} in *all* domains. Instead, it was to show that S_{VN} corresponds to S_{TD} *only in the domain where phenomenological thermodynamics hold*, in all other cases merely *approximating* S_{TD} in large N cases or replacing it altogether (in e.g. single-particle cases). I maintain that **H**&**S**'s main mistake was to confuse the domain where phenomenological thermodynamics hold, with domains where they do not hold.

H&**S** complain that "[...] real physical systems are finite. This means that Von Neumann's argument does not establish a conceptual identity between the Von Neumann entropy and thermodynamic entropy of physical systems. Identities of physical properties mean that the two quantities refer to the same magnitude in the world." (**H**&**S** 2006, 172) In line with what I have said in §2.4.1, it seems that there was no physically meaningful theoretical term in phenomenological thermodynamics that *could* refer to some quantity in the single-particle case, which was why von Neumann needed to come up with a new measure of entropy to begin with. Furthermore, extending a concept to a new domain does not require strict identity, as we have seen and understood for a long time in the case of statistical mechanics and phenomenological thermodynamics.

As Peres (2002) summarizes: "There should be no doubt that von Neumann's entropy. . . is equivalent to the entropy of classical thermodynamics. (This statement must be understood with the same vague meaning as when we say that the quantum notions of energy, momentum, angular momentum, etc., are equivalent to the classical notions bearing the same names)." (Peres 2002, 174) 'Equivalence' here should not be understood in terms of strict (or conceptual) identity i.e. correspondence at all domains. Rather, we should understand equivalence loosely as correspondence in the suitable domains of application, and successful extension of old concepts in these domains to new domains. As Peres noted above, 'equivalence' should be understood in the context of discovery, where one is trying to develop new concepts which are analogous to old ones in different domains. For von Neumann, we have a theory (phenomenological thermodynamics) that is well-understood, but also another theory (quantum mechanics) that we want to understand in light of the former theory. Finding correspondence provides us with ways to *extend* concepts from the original theory to the new theory: for example, with S_{VN} we may now define 'something like' S_{TD} whereas before there was no way to talk about these cases. The same goes for statistical mechanics: by finding a correspondence between e.g. temperature to mean kinetic energy in the thermodynamic limit, we can *extend* the notion of 'something like' temperature beyond its original domain into systems with small numbers of particles, whereas before there was, again, no way to talk about these cases.

I see nothing wrong in these cases in the context of discovery. We should give up a strong and untenable notion of conceptual identity in this context. If so, **H**&**S**'s objection loses much bite.

They further claim that "the fact that the behavior of the two quantities coincides approximately for a very large number of particles is not enough, because in any ensemble of finite gases there are systems in which the identity will not be true. This means that in a real experiment the Von Neumann entropy is not identical with the thermodynamic entropy." (**H**&**S** 2006, 172) This again reveals a confusion between phenomenological and statistical thermodynamics. If they want to talk about particles *at all*, it seems they must adopt some form of statistical mechanical picture with microscopic variables, given phenomenological thermodynamics' emphasis on purely macroscopic variables like volume or temperature. Yet, if so, they must recognize that thermodynamic entropy S_{TD} is in general not strictly identical to statistical mechanical entropy, e.g. the Gibbs entropy or information entropy (briefly discussed in §2.4.2) *either*. Their complaint about approximate coincidences not being enough for (the relevant sort of) equivalence thus weakens significantly, especially since they must assume some such equivalence (which cannot be strict identity) to even talk about particles within the context of phenomenological thermodynamics to begin with. Furthermore, statistical mechanics is evidently empirically successful in explaining and predicting traditionally thermodynamic phenomena despite this 'non-equivalence' – it is not clear why this 'non-equivalence' should matter if, for all practical purposes, statistical mechanics is the conceptual successor of thermodynamics. Of course, if they could come up with a principled reason why approximations should not be allowed *period*, *while* accounting for statistical mechanics' empirical success in accounting for thermodynamic behavior, then this could change. As of now, I see no such argument forthcoming.

2.5 Conclusion, and Some Open Questions

Given the above, I hope to have shown that **H**&**S**'s argument against the correspondence of S_{VN} and S_{TD} – to my knowledge the only one in the philosophical literature – fails to hold in all three cases considered (§2.4.1 – §2.4.4), as a result of their misunderstanding about domains where phenomenological thermodynamics should hold and domains where it should not. This is compounded with misunderstandings about the role of approximations and the relevant interpretation of density matrices and ignorance in quantum mechanics. I conclude that their argument fails on the whole; the correspondence holds for now.

Of course, even if **H**&**S**'s claims were debunked, this does not yet amount to a positive argument for the equivalence between von Neumann entropy and thermodynamic entropy. Even assuming correspondence, correspondence does not entail equivalence. However, the former does provide good *prima facie* reasons to believe the latter, especially given the novel take on correspondence I provided in the end of §2.3: we can accept the correspondence based on thermodynamic considerations about the Second Law and S_{TD} accounting, but *also* based on quantum mechanical considerations about S_{VN} accounting. The correspondence supports a 'two-way street' – equivalence – between S_{TD} and S_{VN}.

While I hope to have conclusively refuted **H**&**S**'s argument, this is but the beginning of further inquiry into questions arising from this supposed correspondence. Amidst the tangle of entropies, there remains much more housekeeping to be done for philosophers of physics.

Acknowledgements

"Does von Neumann Entropy Correspond to Thermodynamic Entropy?" is, in full, a reprint of the material as it appears in *Philosophy of Science* 88 (1), 145-168, 2021. The dissertation author was the sole author of this paper.

Bibliography

Compagner, A. (1989). “Thermodynamics as the continuum limit of statistical mechanics”. In: *American Journal of Physics* 57, pp. 106–117. DOI: doi:10.1119/1.16103.

Deville, Alain and Yannick Deville (2013). “Clarifying the link between von Neumann and thermodynamic entropies”. In: *Eur. Phys. J. H* 38, pp. 57–81. DOI: DOI:10.1140/epjh/e2012-30032-0.

Earman, John and John Norton (1998). “Exorcist XIV: The wrath of Maxwell’s Demon. Part I. From Maxwell to Szilard.” In: *Studies in History and Philosophy of Modern Physics* 29, pp. 435–471.

— (1999). “Exorcist XIV: The wrath of Maxwell’s Demon. Part II. From Szilard to Landauer and beyond.” In: *Studies in History and Philosophy of Modern Physics* 30, pp. 1–40.

Einstein, Albert (1914/1997). “Contributions to Quantum Theory”. In: *The Collected Papers of Albert Einstein, Volume 6 (English). The Berlin Years: Writings, 1914-1917. (English translation supplement) Translated by Alfred Engel.* Princeton University Press: Princeton, pp. 20–26. DOI: Availableat:https://press.princeton.edu/titles/6161.html.

Hemmo, Meir and Orly Shenker (2006). “Von Neumann’s Entropy Does Not Correspond to Thermodynamic Entropy”. In: *Philosophy of Science* 73.2, pp. 153–174. URL: http://www.jstor.org/stable/10.1086/510816.

Henderson, Leah (2003). “The Von Neumann Entropy: A Reply to Shenker”. In: *British Journal for the Philosophy of Science* 54.2, pp. 291–296. URL: http://www.jstor.org/stable/3541968.

Hughes, R. I. G. (1992). *The Structure and Interpretation of Quantum Mechanics.* Cambridge, Mass: Harvard Univ. Press. DOI: https://doi.org/10.2307/2186092.

Maxwell, James Clerk (1878). “Tait’s “Thermodynamics””. In: *Nature* 17, pp. 257–259, 278–280.

Myrvold, Wayne (2011). “Statistical mechanics and thermodynamics: A Maxwellian view”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42.2, pp. 237–243.

Neumann, John von (1955). *Mathematical Foundations of Quantum Mechanics.* Princeton University Press: Princeton.

Norton, John (2017). “Thermodynamically reversible processes in statistical physics.” In: *American Journal of Physics* 85.135, pp. 135–145.

Peres, Asher (1990). “Thermodynamic Constraints on Quantum Axioms”. In: *Complexity, Entropy, and the Physics of Information, The Proceedings Of The Workshop On Complexity, Entropy and the Physics of Information Held May-June, 1989 in Santa Fe, New Mexico, Santa Fe Institute Studies in the Sciences of Complexity, vol. VIII.* Ed. by Zurek, W. H. Westview Press, pp. 345–356.

— (2002). *Quantum Theory: Concepts and Methods (*2^{nd} *Ed.)* Dordrecht: Kluwer Academic Publishers.

Prunkl, Carina E. A. (2020). “On the Equivalence of von Neumann and Thermodynamic Entropy”. In: *Philosophy of Science* 87.2, pp. 262–280. DOI: 10.1086/707565.

Shenker, Orly (1999). “Is $-ktr(\rho log \rho)$ The Entropy in Quantum Mechanics?” In: *British Journal for the Philosophy of Science* 50, pp. 33–48.

Chapter 3

T Falls Apart: On the Status of Classical Temperature in Relativity

> *Dear Laue: I hear the voice of my conscience when I remind you of the dispute concerning the rendering of the fundamental thermodynamic concepts in the special-relativistic form. There is actually no compelling method in the sense that one view would simply be 'correct' and another 'false'. One can only try to undertake the transition as naturally as possible.*
>
> – Albert Einstein,
> 1953 letter to Max von Laue

3.1 Introduction

Do the laws and concepts of classical thermodynamics (**CT**) hold a universal character? Einstein, for instance, wrote that "[**CT**] is the only physical theory of universal content concerning which I am convinced that, within the framework of the applicability of its basic concepts, it will never be overthrown." (1946/1979, 33) Given such proclamations, and how research in black hole thermodynamics – birthed by formal analogies with **CT** – continues to this day, one naturally assumes that **CT** *can* be extended into the relativistic regime and beyond – there is no limit to the "framework of applicability" of its basic concepts.[1]

[1] For more on black hole thermodynamics and its formal analogies with thermodynamics, see e.g. Bekenstein's (1973 and 1975). See Dougherty & Callender (2016) for criticism, and Wallace (2018) for a rejoinder.

It is therefore interesting that a parallel debate drags on without resolution. Although Planck and Einstein pioneered the special relativistic extension of thermodynamical concepts by developing a set of Lorentz transformations, they by no means settled the issue. Importantly, *temperature* resists a canonical relativistic treatment: there are different equivocal ways of relativizing temperature. While physicists appear to treat this as an empirical problem,[2] or something to be settled by convention,[3] the issue seems *conceptually* problematic to me.

I argue that this situation suggests a breakdown of the classical non-relativistic concept of temperature – $\mathrm{T}_{classical}$ – in special-relativistic regimes, i.e. when we consider the temperature of a relatively moving body at high speeds. The procedures which jointly provided physical meaning to $\mathrm{T}_{classical}$ do not do so in relativistic settings. $\mathrm{T}_{classical}$ breaks down in this regime; there *is* a limit to the framework of applicability of the *classical* thermodynamic concepts.

Notably, my argument will rest *not* on the fact that there is no way of defining temperature in relativistic regimes, but that there are *many*, *equally valid* procedures for defining the relativistic temperature which *disagree* with each other. I focus on four procedures:

- one can attempt to construct a relativistic Carnot cycle,
- use a co-moving thermometer,
- consider a relativistic extension of kinetic theory and particle mechanics,
- or scrutinize the black-body radiation of a moving body.

I chose these four because their classical counterparts were significant in determining the physical meaning of $\mathrm{T}_{classical}$: its theoretical relationship with heat (via the Carnot cycle), its phenomenology (with a thermometer), its ontology (via particles), and its connection with radiation (via black-body radiation). It is through this lens that I propose we understand

[2] For instance, Farias et al (2017) remarks that “the long-standing controversy [...] is mainly based on the initial assumptions, which need to be tested [...] to discern which set of Lorentz transformations is correct for quantities such as temperature”.

[3] Landsberg & Johns (1967) suggests that the choice of Lorentz transformation for temperature could be “settled by convention”.

Einstein's above notion of 'natural'-ness: there is strong *consilience* between these procedures, in the operational sense that the temperature established via any of these procedures agrees with the temperature in other procedures.[4] Contrariwise, their relativistic counterparts demonstrate no such consilience: different procedures predict starkly different behaviors for the temperature of a moving body. Furthermore, for each procedure, we find conceptual difficulties too. 'Natural' procedures in **CT** – which generated a consilient and robust concept of temperature – do not appear to be 'natural' at all in relativistic settings.

I end by proposing two possible interpretations of this situation: an eliminativist one on which we interpret temperature akin to simultaneity, or a pluralist one on which we interpret temperature akin to relativistic rotation.

3.2 The Quest to Relativize Thermodynamics: The Odd Case of Temperature

I focus on attempts to relativize **CT**,[5] i.e. some extension of its laws and concepts into *special* relativity.[6]

The pioneers of relativistic thermodynamics, e.g. Einstein (1907) and Planck (1908), sought a set of Lorentz transformations for the laws and quantities of **CT**,[7] just as we have for e.g. position and time. For instance, an observer O' (or the associated inertial frame) with positions and times (x', y', z', t') moving along the x-axis away from another observer O (and their inertial frame) at constant velocity v can be understood by O to be at positions and time (x, y, z, t) via:

[4] Given a proper understanding of the approximation, idealization, and de-idealization procedures.

[5] I refer to the usual classical / phenomenological set of laws governing a system's approach to equilibrium, the meaning of equilibrium, conservation of energy in terms of heat and work, entropy non-decrease, and entropy at the absolute zero of temperature. See e.g. Planck (1945) for a *locus classicus* on the topic.

[6] That is, we assume that events occur on a background Minkowski (flat) spacetime with signature $\{-,+,+,+\}$, where the allowed coordinate frames are inertial frames, that is, frames or observers moving at constant velocity (or zero velocity).

[7] The details are excellently summarized in Liu (1992 and 1994).

$$\begin{aligned} t' &= \gamma(t - \tfrac{vx}{c^2}) \\ x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \end{aligned} \tag{3.1}$$

where $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$ is the Lorentz factor, and c is the speed of light.

Relativistic thermodynamics hopes to find similar transformations for thermodynamic quantities like temperature, pressure, volume, etc. The underlying assumption is that thermodynamics can be shown to have physical meaning in relativistic regimes only when we have a set of Lorentz transformations under which thermodynamic quantities can be shown to transform, just as we do for position and time.[8]

Planck and Einstein successfully derived the transformations for most thermodynamic quantities like pressure p, volume element dV, and entropy S:[9]

$$\begin{aligned} dV' &= \tfrac{dV}{\gamma} \\ p' &= p \\ S' &= S \end{aligned} \tag{3.2}$$

Fixing S appears to indirectly fix the concepts of heat and temperature, via the well-known relation $Q = TdS$. However, surprisingly, the Lorentz transformation for *temperature* turns out to be highly equivocal.

3.3 The Classical Temperature

In **CT**, at least four well-known procedures exist for establishing the concept of temperature. Notably, there is significant *consilience* between them, which suggests there really is

[8] That the only physically meaningful quantities are ones which are invariant or covariant under Lorentz transformations, and that the laws must hold true in similar fashion in all inertial frames, is a common idea in relativity. See Lange (2002, 202) or Maudlin (2011, 32) for an exposition of this idea.

[9] The argument for entropy's Lorentz invariance is generally accepted; I will do the same here. However, see Earman (1986, 177–178) and Haddad (2017, 41 – 42) for criticisms of Planck's argument.

a physically significant quantity: $\mathrm{T}_{classical}$.

3.3.1 The Carnot Cycle

The Carnot cycle is a foundational theoretical concept in **CT** by which we can define absolute temperature in terms of heat.[10] The typical idealized example is an ideal gas acting on a piston in a cylinder (the 'engine') while undergoing reversible processes (see Figure 3.1):

1. The gas receives heat Q_2 from a heat bath at temperature T_2 and isothermally expands, doing work on the surroundings.

2. The cylinder is thermally insulated, and the gas adiabatically expands and continues to do work on the environment, decreasing in temperature to T_1.

3. The gas is isothermally compressed at T_1 at the second heat bath, losing Q_1 to the heat bath.

4. The cylinder is thermally insulated, and the gas is adiabatically compressed as the environment continues to do work on the gas.

5. The cylinder is then brought back to the initial heat bath with T_2.

In such a cycle, a simple relationship between the heat exchange to and from a heat reservoir, and their temperature, can be derived. In a foundational statement of the relationship between heat and temperature in classical thermodynamics, Joule and Thomson (1854/1882) wrote:

> If any substance whatever, subjected to a perfectly reversible cycle of operations, takes in heat only in a locality kept at a uniform temperature, and emits heat only in another locality kept at a uniform temperature, the temperatures of these localities are proportional to the quantities of heat taken in or emitted at them in a complete cycle of operations. [say, a Carnot cycle laid out above] (1854/1882, 394)

[10] For an excellent historical account of Lord Kelvin's definition of the classical temperature via the Carnot cycle, see Chang (2004, Ch. 4).

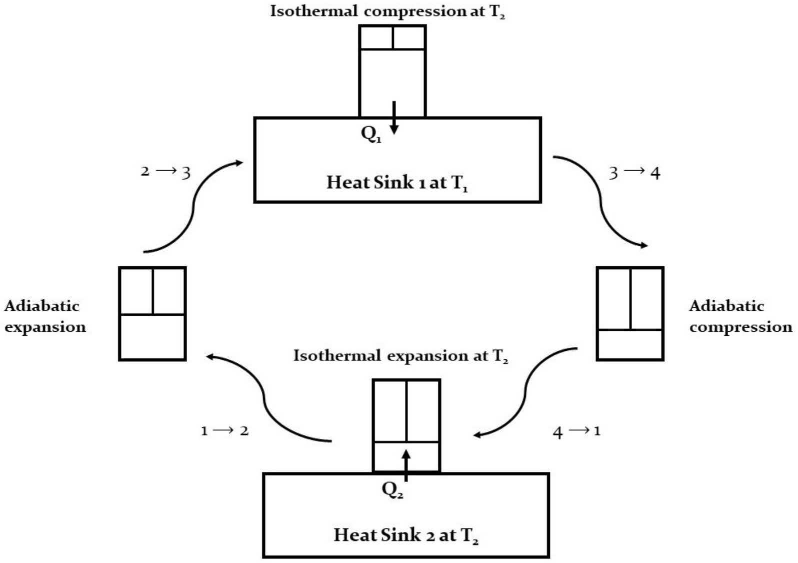

Figure 3.1. An example of a Carnot cycle, with $T_2 > T_1$.

Formally, it is a remarkably simple statement:

$$\frac{T_1}{T_2} = \frac{Q_1}{Q_2} \tag{3.3}$$

If one can calculate the amount of heat exchanged between the two reservoirs, we can then theoretically derive the ratio between the two reservoirs' temperatures. With this in hand, Thomson, as McCaskey (2020, 32) puts it, proposed that "we should construct a temperature scale from the mathematics of a non-existent idealized Carnot engine and the behaviour of a non-existent idealized gas. Any differences between the calculated temperature and the readings of a thermometer will be attributed to shortcomings in the thermometer."

This is the problem: how do we actually connect this to actual measurements and the world? Actual Carnot cycles are hard to construct realistically, and they depend crucially on the fact that gases in these cycles obey the ideal gas law:

$$PV = nRT \tag{3.4}$$

where T is the temperature defined by (3).[11] While no gases are strictly ideal gases, Thomson,[12] and later Callendar (1887) and Le Chatelier & Boudouard (1901), pursued a variety of operationalizations and experiments which succeeded in measuring the extent to which actual gases deviated from the predicted behavior of ideal gases, i.e. deviated from (4): only about $0.628°$ at 1000°C for constant-volume air thermometers, and $1.198°$ for constant-pressure air thermometers. In other words, actual air thermometers approximated theoretical ideal gas thermometers remarkably well. In doing so, we can establish how measurements of temperature derived from observations of actual gases approximates that of the theoretical temperature predicted for ideal gases in Carnot cycles. This, in turn, establishes the legitimacy of the theoretical notion of classical temperature defined here.[13]

3.3.2 The Thermometer

This brings us to thermometers and how they establish the concept of temperature. One crucial development was Fahrenheit's invention of a reliable thermometer, which allowed one to make independent measurements of this physical quantity called the temperature, consistently compare said measurements, and grasp temperature as a robust and (fairly) precisely measured numerical concept rather than one associated with vague bodily sensations. It is not an understatement to say that the classical notion of temperature would not be developed without such an invention.[14]

While thermometers made with the *same* material *were* reliable with respect to each other, thermometers made with *different* materials differed in their rates of expansion and contraction. Importantly, the Carnot cycle discussed above provides a *theoretical* foundation for temperature by providing a definition of temperature independent of material. As Thomson

[11] P is pressure, V is volume, n the amount of substance, and R the ideal gas constant.

[12] For a discussion of Thomson's strategies and the extent to which they succeeded, see Chang & Yi (2005).

[13] For a much more detailed discussion of just how complicated these operationalizations were, and why they are nevertheless successful, see Chang (2004, sub-section "Analysis: Operationalization—Making Contact between Thinking and Doing").

[14] For an involved discussion of this development, see Chang (2004, chs. 1 and 2) For a general overview, see McCaskey (2020).

himself explained:

> As reference is essentially made to a specific body as the standard thermometric substance, we cannot consider that we have arrived at an absolute scale...

In contrast:

> The relation between motive power and heat, as established by Carnot, is such that quantities of heat, and intervals of temperature, are involved as the sole elements in the expression for the amount of mechanical effect to be obtained through the agency of heat; [...], we are thus furnished with a measure for intervals according to which absolute differences of temperature may be estimated. (Thomson 1848/1882, 102)

In other words, the Carnot cycle is intended to provide theoretical foundations to the observed measurements of temperature provided by actual thermometers.

However, as we have already seen, actual thermometers themselves were in turn crucial for supporting the theoretical notion of temperature. Given the difficulties in building an actual Carnot engine and the non-ideal nature of actual gases, actual air thermometers provided a means of de-idealization. They showed how actual temperature measurements can be understood as approximating the abstract theoretical notion of temperature defined by an idealized Carnot cycle and ideal gases.

Put another way, in yet another sign of consilience, we can understand the concept of classical temperature to be co-established by both the observed quantity of temperature provided by a thermometer *and* the theoretical quantity provided by the Carnot cycle, with each procedure supporting the other.

3.3.3 Kinetic Theory of Heat

The kinetic theory of heat provides another way to understand temperature via the Maxwell-Boltzmann distribution. For a system of ideal gas[15] in equilibrium with temperature T, the Maxwell-Boltzmann distribution connects the notion of temperature explicitly with the notion of bulk particle velocities via:

[15] Here an ideal gas is interpreted as a set of n identical weakly interacting particles.

$$f(v) = \sqrt{\left(\frac{m}{2\pi kT}\right)^3} 4\pi v^2 e^{-[\frac{1}{2}mv^2 + V(x)]/kT} \tag{3.5}$$

where m is the particle's mass, T is the temperature, k is Boltzmann's constant, V(x) is the system's position-dependent potential energy, and v is an individual molecule's velocity.[16] Importantly, $f(v)$ tells us, for a system of ideal gas in equilibrium, how many particles we expect to find with some range of velocities v to $v + dv$, given some temperature.

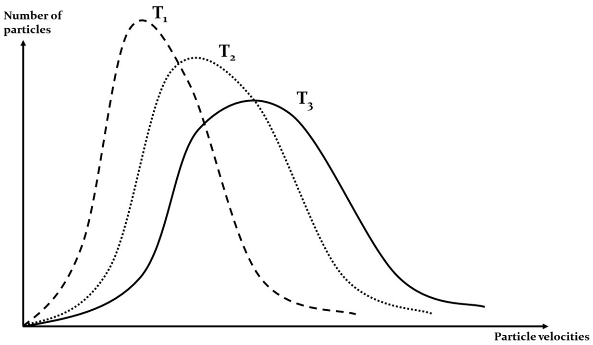

Figure 3.2. A schematic Maxwell-Boltzmann distribution at three different temperatures $T_1 < T_2 < T_3$.

Figure 3.2 shows the schematic connection between a gas's temperature and velocities of the particles composing said gas. This distribution plays a significant conceptual role by allowing us to derive the well-known relationship between temperature and mean kinetic energy:

$$\langle \frac{1}{2}mv^2 \rangle = \frac{3}{2}kT \tag{3.6}$$

This provides foundational support for thermodynamics – and the concept of temperature – in terms of particle mechanics, by allowing us to understand the concept of a system's temperature in terms of the mean kinetic energy of particles composing said system.[17]

Importantly, though, this distribution holds only for ideal gases. As such, the de-

[16] For a historical account of this distribution, see Brush (1983, §1.11).

[17] This has been much discussed in the philosophical literature. See e.g. De Regt (2005) and references therein.

idealization of ideal gases in terms of actual gases, already mentioned in previous sections, also plays a role here and further highlights the consilience of these various procedures in establishing the concept of temperature. They show us that measured temperatures of actual gases are approximately the temperatures of ideal gases with similar pressures and volumes; in turn, the idealized relationship between mean kinetic energy and temperature here, which holds for ideal gases, can also be understood to approximate that of real systems of particles.[18] This provides us with a reason to accept that the temperature of actual systems really can be understood in terms of mean kinetic energy as well, providing further physical meaning to the concept of temperature. There is, again, a consilience between the concept of temperature employed here and the concept of temperature developed through the other procedures.

3.3.4 Black-Body Radiation

Finally, the study of black-body radiation connects temperature to electromagnetic radiation. A black-body is defined as one which absorbs (and emits) all thermal radiation incident upon it without reflecting or transmitting the radiation, for *all* wavelengths and angles of incident upon it. Notably, since a black-body does not distinguish directionality, it emits *isotropic* radiation.

There are simple laws relating the properties of radiation to the black-body's temperature.[19] Firstly, the Stefan-Boltzmann law states:

$$j^* = \sigma T^4 \tag{3.7}$$

where j^* is the total energy emitted per unit surface area per unit time by the black-body, T is its temperature, and σ is the Stefan-Boltzmann constant.[20]

[18] In particular, highly dilute gases at non-extremal temperatures.

[19] For a historical account, see Brush (1983, §3.1) or Stewart & Johnson (2016).

[20] For less idealized bodies which do not absorb all radiation, the law is given by:

$$j^* = \epsilon\sigma T$$

where $0 < \epsilon < 1$ is the emissivity of the substance.

Secondly, Wien's displacement law:

$$\rho(f,T) = f^3 g(\frac{f}{T}) \tag{3.8}$$

states that the energy density ρ of radiation from systems with temperature T, at frequency f, is proportional to $f^3 g(\frac{f}{T})$ for some function g.[21] Integrating over all possible f amounts to computing the total energy density of radiation from all frequencies, and entails the Stefan-Boltzmann law regardless of choice of g. Furthermore, if $\rho(f,T)$ achieves its maximum for some value of f, f_{max}, then:

$$f_{max} \propto T \tag{3.9}$$

or in terms of peak wavelength λ_{peak}:

$$\lambda_{peak} \propto \frac{1}{T} \tag{3.10}$$

This captures the familiar observation that things which are heated first turn red and then into other colors associated with higher frequencies – and hence shorter wavelengths – as their temperature increases.

These laws form the foundations of the relationship between radiation and temperature for a radiating black-body in equilibrium in **CT**. In particular, much of the work in the late 19th century revolved around the search for the appropriate function g, and modifications or generalizations to Wien's law, pursued by others like Rayleigh, Planck, and Einstein.

Interestingly, as another mark of consilience, Einstein's famous 1905 discussion of the quantization of thermal radiation drew upon analogies between thermal radiation and the ideal gas law as well as the Maxwell-Boltzmann distribution for ideal gases: the energy density formula one extracts from the former looks remarkably like the latter.[22] This led Einstein to

[21] See Brush (1983, Ch. 3) for a historical narrative.

[22] See Norton (2005) for a discussion of the analogies and disanalogies between the two. See also Uffink (2006).

conclude that we can understand thermal radiation as quantized, analogous to how an ideal gas can be understood as composed of a number of particles. Crucially for our purposes, we once again see how the same classical temperature concept is supported and applied across different procedures, establishing temperature as a physically significant and meaningful concept across these contexts.

3.4 From Classical to Relativistic Temperature

In **CT**, the above procedures show remarkable consilience in that the very same concept of temperature can be determined or understood in terms of any of these procedures without much practical issue. For instance, the temperature observed by a thermometer for a radiating body is approximately the temperature deduced via the observed frequencies of their radiation. A box of gas can equilibrate with a radiating system and both will come to the same temperature.[23] Finally, the theoretical definition of temperature found by considering a Carnot cycle can also be connected back to the empirical temperature measurements of actual thermometers.

This consilience then motivates why we might find the concept of temperature – and its application in these contexts – 'natural', to borrow Einstein's words. After all, these procedures clearly refer to some quantity which can be measured, manipulated, understood, compared, and calculated across various contexts.

However, there is no such consilience when attempting to relativize temperature. Each procedure establishes a *different* notion of relativistic temperature, and not without conceptual difficulties.

See Fowler (n.d., "Einstein Sees a Gas of Photons") for a preliminary exposition of the analogy.

[23] Einstein's famous 1905 discussion of the quantization of thermal radiation drew upon analogies between thermal radiation and the Maxwell-Boltzmann distribution: the energy density formula one extracts from the former looks remarkably like the latter. This led Einstein to conclude that we can understand thermal radiation as quantized, analogous to how an ideal gas can be understood as composed of a number of particles. See Norton (2005) and Uffink (2006) for discussion. See Fowler (n.d., "Einstein Sees a Gas of Photons") for a preliminary exposition of the analogy.

3.4.1 The Relativistic Carnot Cycle: Moving Temperature is Lower/Higher

I begin with the relativistic Carnot cycle.[24] Von Monsengeil, who devised the process, explicitly appealed to the foundational role of the classical Carnot cycle in defining T$_{classical}$, and proposes an extension to that procedure. (von Mosengeil 1907, 160 – 161) Essentially we demand that the same relations between heat and temperature hold when one heat bath is now *moving* with respect to the other with some velocity v (see Figure 3.3). We are supposed to adiabatically accelerate the engine (i.e. the piston and cylinder of gas) from one inertial frame to another, and adiabatically decelerate it back to the original frame in completing this cycle.

Just as a classical Carnot cycle defined a relationship between the temperature and heat exchange of two heat baths, a relativistic Carnot cycle is stipulated to do the same. For a heat bath at rest with temperature T_0 and another moving with respect to it with a 'moving temperature' T', with the engine *co-moving* with the respective heat baths during the isothermal processes:

$$\frac{T'}{T_0} = \frac{Q'}{Q_0} \tag{3.11}$$

and hence:

$$T' = \frac{Q'}{Q_0} T_0 \tag{3.12}$$

What remains is 'simply' to define the appropriate heat exchange relations. That turns out precisely to be the problem: there are two ways to understand the heat exchange between the engine and the moving heat bath, and there does not seem to be a fact of the matter which is appropriate.[25]

Firstly, one may, like Planck and early Einstein, understand the heat transfer from the

[24] See Liu (1992), Liu (1994) for a detailed historical overview of the topic. See Farias et al (2017) for a physics-oriented overview, and Haddad (2017, 39 – 42) for a concise overview of the disagreements on this procedure.

[25] See Liu (1992) for a much more detailed discussion of this disagreement.

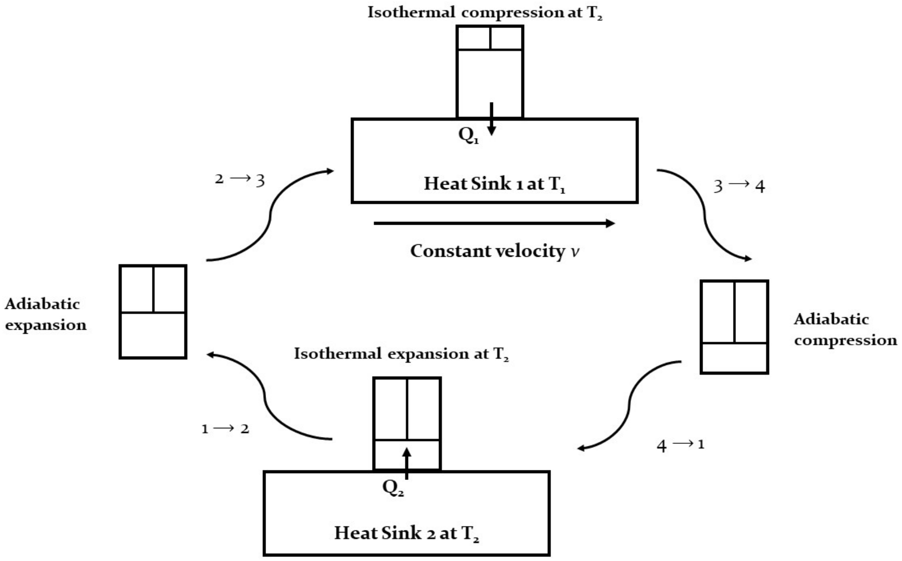

Figure 3.3. A relativistic Carnot cycle.

perspective of the rest frame of the stationary bath. Then the engine, in exchanging heat with the moving bath, also exchanges energy. By relativistic mass-energy equivalence, this causes the bath to lose or gain momentum by changing its mass. However, without further work, the bath then cannot stay in inertial motion - it will decelerate or accelerate. Hence, to keep it moving inertially, we need to perform extra work on it, which Einstein proposed to be:

$$dW = pdV - \mathbf{u} \cdot d\mathbf{G} \tag{3.13}$$

pdV is simply the usual compressional work done by the piston due to the gain or loss of heat from the bath. However, there is a crucial inclusion of the $\mathbf{u} \cdot d\mathbf{G}$ term, where $\mathbf{u}$ is the relativistic velocity of the moving bath (more specifically, $\frac{\mathbf{v}}{c}$) and $d\mathbf{G}$ is the change of momentum due to the exchange of heat. Einstein dubbed this the 'translational work'. One can see that when $\mathbf{u}$ is 0, the work done reduces to the usual definition. In turn, we generalize the first law from:

$$dU = dQ - pdV \tag{3.14}$$

to:

$$dU = dQ - pdV + \mathbf{u} \cdot d\mathbf{G} \tag{3.15}$$

for a moving system. With this definition of work, and some results from continuum electrodynamics, one can obtain a relationship between the quantities of heat exchanged:[26]

$$\frac{dQ'}{dQ_0} = \frac{1}{\gamma} \tag{3.16}$$

and hence, from (11):

$$T' = \frac{1}{\gamma} T_0 \tag{3.17}$$

where γ is the Lorentz factor. We thus arrive at a Lorentz transformation for temperature, according to which a moving system has a lower temperature and appears cooler than a system at rest. This is the *Planck-Einstein formulation* of relativistic thermodynamics.

Secondly, one may, like later Einstein (in private correspondence to von Laue), Ott (1963) and others, doubt the need for translational work. Later Einstein wrote:

> When a heat exchange takes place between a reservoir and a 'machine', both of them are at rest with each other and acceleration-free, it does not require work in this process. This holds independently whether both of them are at rest with respect to the employed coordinate system or in a uniform motion relative to it. (Einstein 1952, quoted in Liu 1994, 199)

In the rest frame of the *moving heat bath*, heat exchange is assumed to occur isothermally (as with the usual Carnot cycle) when both the engine and the heat bath are *at rest* with respect to each other. From this perspective, everything should be as they are classically. There should thus be no additional work required other than that resulting from the heat exchange. Put another way, what was thought of as work done to the system in the Planck-Einstein formulation should instead understood as part of heat exchange in the Einstein-Ott-Arzeliés proposal. Without the translational work term in the equation for work, the moving temperature transformation is

[26] See Liu (1994, 984 – 987) for a full derivation.

instead given by:[27]

$$T' = \gamma T_0 \tag{3.18}$$

and contra (16), the temperature of a moving body appears *hotter*.

I won't pretend to resolve the debate here. However, note that this procedure is not unequivocal on the concept of relativistic temperature: it either appears lower (on the Planck-Einstein formulation) or higher (on the Einstein-Ott formulation) than the rest frame temperature. Importantly, the two proposals reduce to the same classical temperature concept in the rest frame, since the translational work vanishes in this case on both proposals. T$_{classical}$ seems safe, though the fate of its relativistic extension remains undecided.

I end by raising some skepticism about the very idea of a relativistic Carnot cycle, by asking whether there can be a principled answer to whether energy flow is to be understood as 'heat' or 'work' in such a setting. As Haddad observes, this is problematic due to how energy and momentum are interrelated quantities. A system's energy cannot be uniquely decomposed into heat exchange, internal energy and work:

> In relativistic thermodynamics this decomposition is not covariant since heat exchange is accompanied by momentum flow, and hence, there exist nonunique ways in defining heat and work leading to an ambiguity in the Lorentz transformation of thermal energy and temperature. (Haddad 2017, 39)

Since heat flow is accompanied with momentum flow, heat exchange can always be reinterpreted as work done (i.e. as the translational work term).[28] This raises some initial doubts about the very applicability of thermodynamics beyond the rest frame (i.e. **CT** in quotidian settings), given the fundamentality of heat and work relations in thermodynamics.

[27] See Liu (1992, 197 – 198) for a detailed derivation.

[28] See also Dunkel et al (2009, 741).

3.4.2 The Co-Moving Thermometer: 'Moving' Temperature Stays the Same

The idea that relativistic thermodynamics is essentially 'just' quotidian **CT** is echoed by Landsberg (1970), who builds on the classical concept of a thermometer (and the concept of temperature it establishes) via a *co-moving* thermometer:

> One has a box of electronics in both [the relatively moving frame] and [the rest frame] and one arranges, by the operation of buttons and dials to note in [the relatively moving frame] the rest temperature T_0 of the system. This makes temperature invariant. (259)

The co-moving temperature of any system is stipulated to be its relativistic temperature. But this is *no different than the rest frame temperature of that system.* So the Lorentz transformation according to this procedure is simply

$$T' = T_0 \tag{3.19}$$

This proposal can be seen as an extension of T$_{classical}$, in the sense that there is *some* proposed Lorentz transformation. In practice, though, nothing is different from the classical application of a thermometer: we are just measuring the rest frame temperature of the system, as in **CT**. Landsberg partly justifies this with the claim that "nobody in his senses will do a thermodynamic calculation in anything but the rest frame of the system". (Landsberg 1970, 260) On this view, *contrary to the relativistic Carnot cycle*, relativistic temperature transforms *as a scalar*, something found in many relativistic thermodynamics textbooks (e.g. Tolman 1934).

Landsberg (1970, 259) provides an argument for why we couldn't also use this procedure to trivially define alternative Lorentz transformations of other mechanical quantities, e.g. position or time, in terms of their rest frame quantities. He claims that for these mechanical quantities, there *are* measurement discrepancies for the same events in different frames, which needed to be reconciled by Lorentz transformations for consistency. However, for temperature:

> Measurements in a general [reference frame] can be made of mechanical quan-

> tities, but in my view not of temperature, [so] our prescription for T' – namely "measure T_0" – is quite unsuitable for extension to mechanical quantities. (1970, 259)

Prima facie, Landsberg is proposing a novel Lorentz transformation for temperature. However, in my view, this argument amounts to the claim that *there is no relativistic temperature to speak of*; we simply insist on the classical – rest frame – temperature concept. His comparison with mechanical quantities makes this clear: the concept of temperature understood via Landsberg's proposal is *not* relativistic the way other quantities are.

If anything, the preceding discussion suggests that the concept of temperature and its measurement cannot be extended past the rest frame, i.e. into the relativistic domain.[29] As Liu (1994, 992) notes: "The fact seems to be that temperature measurement requires genuine thermal interaction and the state of equilibrium, but when relative macroscopic motion is present, such interaction always disrupts the state of equilibrium and thus renders temperature measurement impossible." Anderson says the same:

> "Thermodynamic quantities only have meaning in the rest frame of the system being observed. [...] This is not to say that an observer could not infer from measurements on a moving system what its rest temperature is. *The point is that he must interpret these measurements in terms of the rest temperature of the system, since this quantity alone depends on thermodynamic state of the system.*" (Anderson 1964, 179 – 180, emphasis mine)

Repeating Landsberg's words in a different context: that "nobody in his senses will do a thermodynamic calculation in anything but the rest frame of the system" suggests that the thermodynamic concept of temperature involved here cannot be extended beyond the classical regime.

[29] This is just what physicists do when they consider the temperature of distant astrophysical bodies. They extrapolate and observe other properties of a body – like luminosity – associated with its *rest frame* temperature. No consideration of moving temperature is involved.

3.4.3 Relativistic Kinetic Theory: No Fact of the Matter

An ideal gas can be understood in terms of particles whose velocities are distributed according to the gas's temperature (§3.3). How does *that* notion of temperature extend to relativistic regimes?

Cubero et al (2007) analyzes the Maxwell-Jüttner distribution, a Maxwell-Boltzmann-type distribution for ideal gases moving at relativistic speeds. They conclude that the temperature should transform as a scalar, i.e. Landsberg's proposal. Interestingly, they explicitly choose a reference frame in which the system is stationary and in equilibrium. But that's just the rest frame of the system! In that case it's unsurprising that there is no transformation required at all for the temperature concept.[30]

Elsewhere, Pathria (1966, 794) proposes yet another construction.[31] They considered a distribution F for an ideal gas in a moving frame with some relativistic velocity $\mathbf{u} = \frac{\mathbf{v}}{c}$:

$$F(\mathbf{p}) = [e^{(E - \mathbf{u} \cdot \mathbf{p} - \mu)/kT} + a]^{-1} \tag{3.20}$$

where $\mathbf{p}$ is a molecule's momentum, E its energy, μ the chemical potential, k the Boltzmann constant, T the system's temperature in that moving frame, and a is 1 or -1 for bosonic and fermionic gases respectively. The distribution then tells us, as with the classical case, how many particles we expect to see with momentum $\mathbf{p}$. F is shown to be Lorentz-invariant, and we can compare them as such:

$$\frac{E - \mathbf{u} \cdot \mathbf{p} - \mu}{kT} = \frac{E_0 - \mu_0}{kT_0} \tag{3.21}$$

With the (known) Lorentz transformations for energy and momentum, we can then show that $T = \frac{1}{\gamma} T_0$, i.e. the *Planck-Einstein* formulation.

[30] Cubero et al (2007, 3) admits as much when they note that "Any (relativistic or nonrelativistic) Boltzmann-type equation that gives rise to a universal stationary velocity PDF implicitly assumes the presence of a spatial confinement, thus *singling out a preferred frame of reference.*"

[31] My presentation follows Liu (1994).

One might think that this suggests some consilience between the kinetic theory and the relativistic Carnot cycle *for* the Planck-Einstein formulation. However, one would be disappointed. Balescu (1968) showed that Parthria's proposed distrbution (20) can be generalized as:

$$F^*(\mathbf{p}) = [e^{\alpha(\mathbf{u})(E - \mathbf{u}\cdot\mathbf{p} - \frac{\mu}{\beta(\mathbf{u})})/kT} + a]^{-1} \tag{3.22}$$

with the only constraint that $\alpha(\mathbf{0})$ and $\beta(\mathbf{0})$ = 1 for arbitrary even functions α and β. F^* tells us the particle number (or, in quantum mechanical terms, occupation number) associated with some $\mathbf{p}$ or E over an interval of time. Balescu shows that any such distribution recovers the usual Maxwell-Boltzmann-type statistics, in the sense that distributions with arbitrary choices of these functions all *agree* on the internal energy and momenta in the rest frame when $\mathbf{u} = 0$: T$_{classical}$ is safe from these concerns.

Choosing these functions amounts to choosing some velocity-dependent scaling for temperature via α and chemical potential via β. Importantly, the question of how temperature scales when moving relativistically is precisely what we want to decide on, yet it is also the quantity rendered arbitrary by this generalization! In particular, Balescu shows that:

1. The choice α = 1 amounts to choosing the Planck-Einstein formulation $T = \frac{1}{\gamma}T_0$,
2. The choice α = γ^2 amounts to choosing the Einstein-Ott-Arzeliés formulation $T = \gamma T_0$,
3. The choice $\alpha = \gamma$ amounts to choosing Landsberg's formulation $T = T_0$.

As Balescu notes: "Within strict equilibrium thermodynamics, there remains an arbitrariness in comparing the systems of units used by different Lorentz observers in measuring free energy and temperature" and that "equilibrium statistical mechanics cannot by itself give a unique answer in the present state of development." (1968, 331) Any such choice will be a *postulate*, not something to be assured by the statistical considerations here. In other words, contrary to the classical Maxwell-Boltzmann case, there appears to be no fact of the matter how temperature will behave relativistically, given the underlying particle mechanics.

Contrary to the classical kinetic theory of heat, which provided a unequivocal conceptual picture (and putative reduction) of T$_{classical}$, there's again no such univocality here.

3.4.4 Black-Body Radiation: No Thermality for Moving Black-Bodies

Finally, when we consider *moving* black-bodies, there is again no clear verdict on the Lorentz transformation for the relativistic temperature. The very concept of a black-body appears to be restricted to the rest frame.

McDonald (2020) provides a simple example of why this is so: consider some observed Planckian (thermal) spectrum of wavelengths from some distant astrophysical object with a peak wavelength λ_{peak}. We want to ascribe some temperature to that object directly. In our rest frame, using Wien's law (9):

$$\lambda_{peak} = \frac{b}{T} \tag{3.23}$$

where b is Wien's displacement constant. Supposing we know the velocity **v** of the distant astrophysical object, we can compare wavelengths over distances in relativity using the relativistic Doppler effect to find the peak wavelength of the object λ'_{peak} at the source:

$$\lambda'_{peak} = \frac{\lambda_{peak}}{\gamma(1 + \frac{vcos\theta}{c})} \tag{3.24}$$

where θ is the angle in the rest frame of the observer between the direction of **v** and the line of sight between the observer and the object. Given this, we can compare temperatures:

$$T' = \frac{\lambda_{peak}}{\lambda'_{peak}} T = \gamma(1 + \frac{vcos\theta}{c})T \tag{3.25}$$

The predicted temperature thus depends on the *direction* of the moving black-body to the inertial observer.

Landsberg $\&$ Matsas (1996) shows similar results and demonstrates how a relatively moving black-body generally does not have a black-body spectrum from the perspective of an inertial observer. Crucially, they emphasize just how problematic this is for the notion of *black-body radiation* which is defined as *isotropic*:

> [the equation for a moving black-body] cannot be associated with a legitimate thermal bath (which is necessarily isotropic) [...] the temperature concept of a black body is unavoidably associated with the Planckian thermal spectrum, and because a bath which is thermal in an inertial frame S is non-thermal in [a relatively moving] inertial frame S', which moves with some velocity $v \neq 0$ with respect to S, a universal relativistic temperature transformation [...] cannot exist. (1996, 402–403)

In a follow-up article, they further emphasize that "a moving observer in a heat reservoir can therefore not detect a black-body spectrum, and hence cannot find a parameter which can be identified as *temperature*." (2004, 93)

The general lesson is simple yet profound. A black-body was defined in the rest frame, i.e. in the non-relativistic setting: we see isotropic radiation with a spectrum, which can be understood to be in equilibrium with other objects and measured as such with thermometers. However, there was no guarantee that a *moving* black-body would still be observed as possessing some *black-body* spectrum with which to ascribe temperature. And it turns out that it generally does *not*. Without this assurance, we cannot reliably use the classical theory of black-body radiation to find a relativistic generalization of temperature.

3.5 T$_{classical}$ Falls Apart. What Then?

Examining four relativistic counterparts to classical procedures thus reveals a *discordant* concept: a moving body may appear to be cooler, or hotter, the same, or may not even appear to be thermal at all. Despite how well these procedures worked classically, they do *not* work together to establish a unequivocal concept of relativistic temperature. Furthermore, *within* each procedure, various conceptual difficulties suggests that the concept of relativistic temperature does not find firm footing *either.* Returning to Einstein's quote, it appears that there is no 'natural' way to extend T$_{classical}$.

T$_{classical}$ thus fails to be extended to relativity: well-understood procedures that unequivocally establish its physical meaning in classical settings fail to do so in relativistic settings. These procedures appear to work *just fine* in classical settings, i.e. in the rest frame. However, attempting to extend them to relativistic settings immediately led to conceptual difficulties. This all suggests that the concept of temperature – and correspondingly, heat – is inherently a concept restricted to the rest frame.

More generally, any relativistic extension of **CT** violates some classical intuitions and will appear 'unnatural'. No matter our choice of temperature transformation, something from **CT** must go. Broadening Balescu's point ($\S 4.3$), Landsberg (1970, 263–265) generalizes the thermodynamic relations in terms of arbitrary functions $\theta(\gamma)$ and $f(\gamma)$:

$$TdS = \theta dQ \tag{3.26}$$

$$dQ = f dQ_0 \tag{3.27}$$

where f is the force function:

$$f = \frac{1}{\gamma} + r(1 - \frac{1}{\gamma^2}) \tag{3.28}$$

where $r = 0$ if we demand the Planck-Einstein translational work, or $r = \gamma$ for the Einstein-Ott view without such work. Again, we only require $\theta(1) = f(1) = 1$ so that in the rest frame everything reduces to **CT**. Different choices, again, entail different concepts of relativistic temperature, but also other thermodynamic relations (and hence the thermodynamic laws). (See Figure 3.4.)

Importantly, no choice preserves all intuitions about T$_{classical}$ and **CT**. Demanding that a lower (higher) moving temperature leads to non-classical behavior. Landsberg (1970, 260 – 262) considers two thermally interacting bodies A and B moving relatively to one another. A (in its rest frame) sees the other as cooler (warmer) and hence heat flows from (to) B. But the same analysis occurs in B's rest frame to opposite effect! So heat flow becomes frame-dependent and indeterminate, contrary to our classical intuitions.

Moving temperature...	$dQ/TdS = \theta$	$dQ/dQ_0 = f$	r	$T/T_0 = \theta f$
... is lower	1	$1/\gamma$	0	$1/\gamma$
... is higher	1	γ	γ	γ
... is invariant	γ	$1/\gamma$	0	1

Figure 3.4. A list of some choices of θ, f, and r.

However, demanding temperature-invariance entails that the classical laws of thermodynamics are no longer form-invariant in all inertial frames. Notably, we must revise their form by including some variations of functions f, γ and θ.[32] So we preserve some intuitions about heat flow but give up the cherished form of classical thermodynamical laws. Interestingly, it is precisely this classical form that Bekenstein (1973) appealed to, when making the formal analogies between thermodynamics and black holes.

What then? I end with two possible interpretations of my analysis: an *eliminativist* viewpoint, and a *pluralist* viewpoint.[33] On the former, one might interpret temperature akin to *simultaneity*: both concepts are well-defined within some rest frame, but there is no absolute fact of the matter as to how they apply *beyond* for relatively moving observers. If one believes that the only physically significant quantities are those which are frame-invariant or co-variant (recall fn. 8), temperature's frame-dependence might lead one to abandon talk of temperature as physically significant, just as we have for simultaneity.[34]

On the latter, one might *instead* interpret temperature akin to relativistic *rotation*. Analogously, Malament (2000) identifies two equally plausible criteria for defining rotation

[32] See Landsberg (1970, 264).

[33] See Taylor and Vickers (2017) for discussion of this dichotomy.

[34] For discussion of the status of simultaneity, see Janis (2018) and references therein.

which agree in classical settings, yet disagree in general-relativistic settings. Importantly, both violate some classical intuitions. Nevertheless:

> There is no suggestion here that [this] poses a deep interpretive problem [...] nor that we have to give up talk about rotation in general relativity. The point is just that [...] we may have to disambiguate different criteria of rotation, and [...] that they all leave our classical intuitions far behind. (2000, 28)

Likewise, on this view, we might accept that T$_{classical}$ breaks down, and that (relativistic extensions of) classical procedures fail to unequivocally define a relativistic temperature. However, we need not abandon *temperature* altogether; instead, we need only to work harder to disambiguate and generalize the concept of temperature (and thermodynamical laws).

Depending on interpretation, questions arise. For instance, should formal analogies between black holes and classical thermodynamical laws be taken seriously, if the form of the classical thermodynamical laws doesn't actually survive in relativistic domains? Could typical black holes be treated as 'at rest', such that T$_{classical}$ might still apply? Should we, and how should we, generalize T$_{classical}$? I leave these questions to future work.

3.6 Conclusion

The conclusion that classical thermodynamical concepts fall apart in new regimes should not be surprising to philosophers of science. For instance, Callender (2001) cautioned against "taking thermodynamics too seriously" even in the statistical mechanical regime. He argues that taking the laws and concepts of classical thermodynamics too literally when attempting to formulate a reduction of classical thermodynamics to statistical mechanics leads to error. Furthermore, Callender briefly notes how, similarly, "perhaps the principal reason for the confusion [in relativizing temperature] is the fact that investigators simply assumed that relativistic counterparts of some laws of thermodynamics would look just like the phenomenological laws – they took (some) thermodynamics too seriously." (2001, 551)

Earman (1978, 178) says essentially the same when he diagnoses the problem with relativistic thermodynamics: the pioneers of relativistic thermodynamics acted "as if thermodynamics were a self- contained subject, existing independently of any statistical mechanical interpretation. Within this setting, many different 'transformation laws' for the thermodynamical quantities are possible." However, the problem is somewhat worse if what I've said in §4.3 is right: even the statistical mechanical determination of a relativistic temperature is up for grabs.

In any case, I hope to have highlighted how *messy* the situation is in relativistic thermodynamics. Yet, while physicists continue to chime in,[35] not much has been said by contemporary philosophers, despite "how rich a mine this area is for philosophy of science". (Earman 1978, 157) Besides Earman, the only other philosopher to have discussed this topic in detail appears to be his student, Liu. (1992 and 1994) Through this paper, I hope to have at least re-ignited some interest in this topic.

Acknowledgements

"T Falls Apart: On the Status of Classical Temperature in Relativity", is, in full, a reprint of the material as it is forthcoming in *Philosophy of Science* 90 (5), 2023. The dissertation author was the sole author of this paper.

[35] See McDonald (2020) and references therein.

Bibliography

Balescu, R. (1968). "Relativistic statistical thermodynamics". In: *Physica* 40.3, pp. 309–338. ISSN: 0031-8914. DOI: https://doi.org/10.1016/0031-8914(68)90132-8. URL: https://www.sciencedirect.com/science/article/pii/0031891468901328.

Bekenstein, Jacob D. (Nov. 1975). "Statistical black-hole thermodynamics". In: *Phys. Rev. D* 12 (10), pp. 3077–3085. DOI: 10.1103/PhysRevD.12.3077. URL: https://link.aps.org/doi/10.1103/PhysRevD.12.3077.

Brush, Stephen (1983). *Statistical Physics and the Atomic Theory of Matter From Boyle and Newton to Landau and Onsager.* Princeton: Princeton University Press.

Callendar, Hugh Longbourne (1887). "On the Practical Measurement of Temperature: Experiments Made at the Cavendish Laboratory, Cambridge." In: *Philosophical Transactions of the Royal Society of London* A178, pp. 161–230.

Chang, Hasok (2004). *Inventing Temperature: Measurement and Scientific Progress.* Oxford University Press.

Cubero, David et al. (2007). "Thermal Equilibrium and Statistical Thermometers in Special Relativity". In: *Phys. Rev. Lett.* 99 (17), p. 170601. DOI: 10.1103/PhysRevLett.99.170601. URL: https://link.aps.org/doi/10.1103/PhysRevLett.99.170601.

De Regt, Henk W. (2005). "Kinetic Theory". In: *The Philosophy of Science: An Encyclopedia.* Ed. by Sahotra Sarkar and Jessica Pfeifer. Routledge.

Dunkel, Jörn, Peter Hänggi, and Stefan Hilbert (Oct. 2009). "Non-local observables and lightcone-averaging in relativistic thermodynamics". In: *Nature Physics* 5.10, pp. 741–747. ISSN: 1745-2481. DOI: 10.1038/nphys1395. URL: https://doi.org/10.1038/nphys1395.

Earman, John (1978). “Combining Statistical-Thermodynamics and Relativity Theory: Methodological and Foundations Problems”. In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1978, pp. 157–185.

Einstein, Albert (1907). “Uber das Relativitätsprinzip und die aus demselben gezogenen”. In: *Folgerungen. J. Radioakt. Elektron* 4, pp. 411–462.

— (1946/1979). *Autobiographical Notes. (P. A. Schilpp, Trans.)* Open Court Printing.

Farías, Cristian, Victor A. Pinto, and Pablo S. Moya (2017). “What is the temperature of a moving body?” In: *Nature Scientific Reports* 7.1, p. 17657. ISSN: 2045-2322. DOI: 10.1038/s41598-017-17526-4. URL: https://doi.org/10.1038/s41598-017-17526-4.

Fowler, Michael (n.d.). *Black-Body Radiation.* URL = https://galileo.phys.virginia.edu/classes/252/black_body_radiation.html (last accessed March 19 2022).

Haddad, Wassim M. (2017). “Thermodynamics: The Unique Universal Science”. In: *Entropy* 19.11. ISSN: 1099-4300. URL: https://www.mdpi.com/1099-4300/19/11/621.

Janis, Allen (2018). “Conventionality of Simultaneity”. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Fall 2018. Metaphysics Research Lab, Stanford University.

Joule, James Prescott and William Thomson (1854/1882). *On the Thermal Effects of Fluids in Motion, Part 2.* In Thomson 1882, 357–400. Originally published in the Philosophical Transactions of the Royal Society of London 144:321–364.

Landsberg, P. T . (1970). “Special Relativistic Thermodynamics – A Review.” In: *Critical Review of Thermodynamics.* Ed. by E. B. Stuart, B. Gal-Or, and A. T. Brainard. Baltimore: Mono Book Corp., pp. 253–72.

Landsberg, P. T. and K. A. Johns (1967). “A relativistic generalization of thermodynamics”. In: *Il Nuovo Cimento B (1965-1970)* 52.1, pp. 28–44. DOI: 10.1007/BF02710651.

Landsberg, P. T. and G. E. A. Matsas (2004). “The impossibility of a universal relativistic temperature transformation”. In: *Physica A: Statistical Mechanics and its Applications*

340.1, pp. 92–94. issn: 0378-4371. url: https://www.sciencedirect.com/science/article/pii/S0378437104004017.

Landsberg, Peter T. and George E.A. Matsas (1996). “Laying the ghost of the relativistic temperature transformation”. In: *Physics Letters A* 223.6, pp. 401–403. issn: 0375-9601. doi: 10.1016/s0375-9601(96)00791-8. url: http://dx.doi.org/10.1016/S0375-9601(96)00791-8.

Lange, Marc (2002). *An Introduction to the Philosophy of Physics: Locality, Fields, Energy, and Mass.* Blackwell.

Le Chatelier, Henri and O. Boudouard (1901). *High-Temperature Measurements (Trans. George K. Burgess.* New York: Wiley.

Liu, Chuang (1992). “Einstein and Relativistic Thermodynamics in 1952: A Historical and Critical Study of a Strange Episode in the History of Modern Physics”. In: *British Journal for the History of Science* 25.2, pp. 185–206. doi: 10.1017/s0007087400028764.

— (1994). “Is There a Relativistic Thermodynamics? A Case Study of the Meaning of Special Relativity”. In: *Studies in History and Philosophy of Science Part A* 25.6, pp. 983–1004. doi: 10.1016/0039-3681(94)90073-6.

Maudlin, Tim (2011). *Quantum Non-Locality and Relativity.* John Wiley & Sons, Ltd.

McCaskey, John P. (2020). “History of ‘temperature’: maturation of a measurement concept”. In: *Annals of Science* 77.4, pp. 399–444.

McDonald, Kirk (2020). *Temperature and Special Relativity.* (last accessed March 15 2022). url: http://kirkmcd.princeton.edu/examples/temperature_rel.pdf.

Norton, John D. (2006). “Atoms, Entropy, Quanta: Einstein’s Miraculous Argument of 1905”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 37.1, pp. 71–100. doi: 10.1016/j.shpsb.2005.07.003.

Ott, H. (1963). “Lorentz-Transformation der Warme und der Temperatur”. In: *Zeitschrift fur Physik* 175, pp. 70–104.

Pathria, R K (1966). "Lorentz transformation of thermodynamic quantities". In: *Proceedings of the Physical Society* 88.4, pp. 791–799. DOI: 10.1088/0370-1328/88/4/301. URL: https://doi.org/10.1088/0370-1328/88/4/301.

— (1967). "Lorentz transformation of thermodynamic quantities: II". In: *Proceedings of the Physical Society* 91.1, pp. 1–7. DOI: 10.1088/0370-1328/91/1/302. URL: https://doi.org/10.1088/0370-1328/91/1/302.

Planck, M., Verband Deutscher Physikalischer Gesellschaften, and Max-Planck-Gesellschaft zur Förderung der Wissenschaften (1958). *Physikalische Abhandlungen und Vorträge: Aus Anlass seines 100. Geburtstages (23. April 1958).* Physikalische Abhandlungen und Vorträge. F. Vieweg.

Planck, Max (1908). "Zur Dynamik bewegter Systeme". In: *Ann. Phys. Leipz.* 26, pp. 1–34.

— (1945). *Treatise on Thermodynamics (A. Ogg Trans.)* Dover Publications.

Stewart, Seán M. and R. Barry Johnson (2016). *Blackbody Radiation: A History of Thermal Radiation Computational Aids and Numerical Methods.* CRC Press. DOI: 10.1201/9781315372082. URL: https://doi.org/10.1201/9781315372082.

Taylor, Henry and Peter Vickers (2017). "Conceptual Fragmentation and the Rise of Eliminativism". In: *European Journal for Philosophy of Science* 7.1, pp. 17–40. DOI: 10.1007/s13194-016-0136-2.

Thomson, William (1848). "On an Absolute Thermometric Scale Founded on Carnot's Theory of the Motive Power of Heat, and Calculated from Regnault's Observations." In: In Thomson (1882), 100–106. Originally published in the Proceedings of the Cambridge Philosophical Society 1:66–71; also in the Philosophical Magazine, 3rd ser., 33:313–317.

— (1882). *Mathematical and Physical Papers. Vol. 1.* Cambridge: Cambridge University Press.

Tolman, Richard C. (1934). *Relativity Thermodynamics and Cosmology.* Oxford: Clarendon Press.

Uffink, Jos (2006). “Insuperable Difficulties: Einstein’s Statistical Road to Molecular Physics”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 37.1, pp. 36–70. doi: 10.1016/j.shpsb.2005.07.004.

von Mosengeil, Karl (1907). “Theorie der stationaren Strahlung in einem gleichformig bewegten Hohlraum”. In: *Ann Phys.* 22. Reprinted in Planck (1958), Vol. II, 138–75., pp. 876–904.

Wallace, David (2018). “The case for black hole thermodynamics part I: Phenomenological thermodynamics”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 64, pp. 52–67. issn: 1355-2198. doi: https://doi.org/10.1016/j.shpsb.2018.05.002. url: https://www.sciencedirect.com/science/article/pii/S1355219817301661.

Chapter 4

Do Black Holes Evaporate? The Case of Quasi-Stationarity

4.1 Introduction

It is often said that black holes are where our best theory of matter, quantum mechanics, meet our best theory of spacetime, general relativity. One prominent way in which black hole physics connects both quantum mechanics and general relativity is through the study of black hole thermodynamics. In Curiel's words, "almost everyone agrees that black hole thermodynamics provides our best guide for clues to a successful theory of quantum gravity." (2019, 27) The hope is that a closer investigation of black holes will unearth a more fundamental theory unifying both quantum mechanics and general relativity.[1]

Crucial to the study of black hole thermodynamics is the use of *idealizations* in order to drastically simplify our calculations and render our models of black holes tractable. Here, I want to focus particularly on the idealization of *quasi-stationarity*, an idealization that is widely used by physicists including those studying black holes. As Page (2005, 10) points out, for instance, proofs of the Generalized Second Law generally assume that the black hole in question is assumed to be quasi-stationary, changing only slowly during its interaction with an environment: it is almost unchanging, hence "quasi" and "stationary". This allows us to model the *dynamically* changing black hole with a temporal sequence of *stationary* black holes, each

[1] See e.g. Hawking 1977.

differentiated by different values of some parameter (such as mass). Crucially, the stationarity of the black hole at each point of the sequence allows us to ascribe a *globally conserved energy* to the black hole at all times,[2] something that is importantly generically *not available in general relativity.*[3] However, the use of this idealization is not always made explicit; it might be masked by describing the black hole as ’slowly evolving’ (Zurek & Thorne 1985), or by discussing black hole *dynamics* while using a *static* or *stationary* metric like the Schwarzschild or Kerr metrics. (Abramovicz and Fragile 2013)

Prominently, Hawking made explicit use of quasi-stationarity in his 1975 argument for black hole evaporation. By adopting a semiclassical approximation, Hawking studied the effects of (classical) black hole horizons on a (quantum) vacuum field and argued that black hole horizons radiate at a certain temperature to observers at infinity (i.e. Hawking radiation). This led him to conclude that black holes must evaporate. At a conceptual level, his argument goes something like this: since we observe black holes radiating energy via Hawking radiation, this energy must be coming from somewhere due to the conservation of energy. Since the spacetime in question is vacuum everywhere,[4] the only object that could lose energy as a result of the radiation is the object of study, the black hole. This led Hawking to conclude that black holes must lose mass (‘evaporate’) as a result of Hawking radiation, just as ordinary matter radiate and lose energy to their environment. This is one reason why many now treat black holes as bona fide thermodynamic objects, just like ordinary matter.[5] Crucially, his argument relies on the idealization of quasi-stationarity: dynamical black holes are modeled as a sequence of stationary time-independent black holes, so that black holes could dynamically change – lose mass and hence energy – *over time* while having a conserved energy at every point in time.

[2] Due to the myriad literature out there with differing terminologies, and the mass-energy equivalence in general relativity, I will use ’mass’, ’energy’, and ‘mass-energy’ somewhat interchangeably in this paper.

[3] See Maudlin, Okon and Sudarsky (2020) for an excellent discussion.

[4] More precisely, the spacetime contains a quantum field in the vacuum state.

[5] For a general rejoinder to black hole thermodynamics as more than a ’formal analogy’, see Dougherty and Callender (2016). For a brief defense, see Wallace (2018).

Beyond its ubiquity in the study of black holes, what also interests me is the observation that quasi-stationarity shares much in common with another widely used idealization elsewhere in physics: the quasi-static (or thermodynamically reversible) process in thermodynamics. Just like quasi-stationarity, a system undergoing a quasi-static process is described as evolving so slowly that we can model the system as with some sequence of equilibrium (time-independent) states, each with a different thermodynamic parameter such as temperature, pressure and so on.

Importantly, Norton (2016) has recently pointed out that the quasi-static process, taken literally, is impossible. It requires us to envision a system undergoing dynamical changes *over time*, while being in a time-independent equilibrium state at *all times*. This being contradictory, the quasi-static process cannot literally describe systems in the real world. Nevertheless, Norton shows how we can justify the use of quasi-staticity by providing a de-idealization procedure:[6] We can understand a quasi-static process as the limit of actual irreversible processes containing *non*-equilibrium states. Real systems, in a very concrete sense, approximate the quasi-static process. Because of the existence of such a story, we are justified in continuing our use of the quasi-static process in thermodynamics despite its literally contradictory nature.

In my view, the quasi-stationary process, taken literally, is likewise impossible. I argue that this gives us reason to revisit its ubiquitous use, by arguing that at least one prominent case of its use remains unjustified. I claim that the justificatory strategy of de-idealization does not apply to Hawking's use of quasi-stationarity in his 1975 argument for black hole evaporation. Hawking's argument crucially relies on the limit property of quasi-stationary processes: it requires the system to have a globally conserved energy as a result of being time-independent at all times, while simultaneously requiring the system to undergo dynamical changes. Since we expect the black hole in question to be dynamical, i.e. time-dependent, global conservation of energy cannot literally obtain. For something like Hawking's argument to go through, we need something like an *approximately* globally conserved energy, just as

[6] See e.g. McMullin (1985) for a classic account of de-idealization.

thermodynamic systems undergo approximately quasi-static processes without undergoing a literally quasi-static process. In general relativity, the global conservation of energy depends on the existence of time-like Killing fields, and a de-idealization of this property must thus show how a real system can approximately possess a globally conserved energy by showing the existence of *approximately* time-like Killing fields (as I will explain). However, crucially, I argue that there is no suitable 'de-idealization' procedure available, contrary to the case of quasi-staticity in thermodynamics. There is no clear and unproblematic account of what approximately time-like Killing fields amount to. Until a clear 'de-idealization' procedure can be produced, Hawking's argument remains unjustified.

More generally, to my knowledge, these aspects of quasi-stationarity vis-a-vis its use in black hole physics have not been studied in detail by either philosophers of physics working on black hole thermodynamics or philosophers of science working on the use of idealizations.[7] I thus hope that this paper can fill both lacunae and provide a bridge between general philosophy of science and philosophy of physics, by highlighting how discussions about idealizations in philosophy of science might help us evaluate arguments in physics.

Astute readers might immediately worry that contemporary arguments for black hole evaporation do not necessarily rely on quasi-stationarity. For instance, they might instead rely on something like asymptotic flatness.[8] In a follow-up paper I hope to scrutinize asymptotic flatness and argue that a similar problem obtains for that idealization as well. In short: the problem for approximate Killing fields raised here becomes a problem for approximate *asymptotic Killing fields* in that context of asymptotic flatness. Readers who are (rightfully) concerned about such a worry are urged to read this paper not as a *general* critique of black hole evaporation, but instead as a philosophical analysis of the idealization of quasi-stationarity and one *specific* argument for black hole evaporation for which the use of quasi-stationarity might be unjustified.

[7] One notable exception is Duerr (2019), who criticizes the use of the idealization of asymptotic flatness for justifying the reality of energy in general relativity.

[8] This is Wallace's (2018) presentation of the argument for black hole evaporation.

4.2 Stationarity in General Relativity

I begin by specifying some key concepts which are required for us to more precisely understand the sense of 'stationarity' (and hence 'quasi-stationarity') relevant for my argument here, as well as the connection between stationarity and the global conservation of energy, crucial for Hawking's argument.

To begin with, since work on black hole thermodynamics typically takes place in the semiclassical regime, gravity is understood classically. The arena of discourse is general relativity. Our discussion thus begins from the metric tensor $g_{\alpha\beta}$ (henceforth simply the 'metric'), which defines a spacetime of interest and constrains the behavior of matter on said spacetime via the Einstein Field Equations:[9]

$$R_{\alpha\beta} - \frac{1}{2}Rg_{\alpha\beta} + \Lambda g_{\alpha\beta} = 8\pi T_{\alpha\beta} \tag{4.1}$$

where $R_{\alpha\beta}$ is the Ricci tensor, R is the Ricci scalar, Λ is the cosmological constant (which may or may not vanish), and $T_{\alpha\beta}$ is the stress-energy tensor encoding the behavior of matter in spacetime.

In general, energy is always *locally* conserved in general relativity. Along any worldline χ, the covariant derivative of $T_{\alpha\beta}$ vanishes:

$$\nabla_{\chi} T_{\alpha\beta} = 0 \tag{4.2}$$

In other words, momentum and energy are conserved given infinitesimal displacements along any worldline, as one would expect from classical physics.

However, as is well-known, this does not generally entail *global* conservation of energy.[10] The fact that energy is conserved along any observer's worldline does not allow us to say that

[9] More precisely a spacetime is given by the pair $(M, g_{\alpha\beta})$ where M is a manifold. Here I will speak of the metric and spacetime interchangeably. Nothing turns on this difference. Furthermore, to simplify presentation, I will use natural units such that $c = G = \hbar = k = 1$.

[10] See Maudlin, Okon and Sudarsky (2020) for an excellent discussion.

energy is conserved for the *entire* spacetime.

Famously, Noether (1918) showed that every differentiable symmetry of the action of a physical system is associated with some conserved current satisfying a continuity equation, and thus a corresponding conservation law. In the context of general relativity, these symmetries are represented by Killing vector fields (or simply Killing fields) which generate isometries (trajectories along which the metric is constant). A Killing vector ξ represent an infinitesimal displacement along which the Lie-derivative ($\pounds$) of the metric vanishes:

$$\pounds_{\xi} g_{\alpha\beta} = 0 \tag{4.3}$$

This demand leads naturally to the result that ξ satisfies Killing's equation:

$$\nabla_{\nu}\xi_{\mu} + \nabla_{\mu}\xi_{\nu} = 0 \tag{4.4}$$

With Killing's equation and the geodesic equation, where $\mathbf{p}$ is the tangent vector,

$$\nabla_{\mathbf{p}}\mathbf{p} = 0 \tag{4.5}$$

we can derive the following theorem.[11] In any spacetime geometry endowed with a symmetry described by a Killing field ξ, motion along any geodesic leaves the scalar product of the tangent vector $\mathbf{p}$ with the Killing vector ξ constant:

$$\mathbf{p} \cdot \xi = constant \tag{4.6}$$

This allows us to describe a *globally* conserved quantity on a spacetime.[12] For example, space-like translational symmetries are what allow us to make sense of the global conservation of linear momentum, while the space-like rotational symmetries let us define the global conser-

[11] See Misner, Thorne and Wheeler (1973, 651) for discussion.

[12] See Hawking and Ellis (1973, 61 – 63), Misner, Thorne and Wheeler (1973, §25.2), Carroll (2019, 120), or Maudlin, Okon and Sudarsky (2020, §2.4) for discussion.

vation of angular momentum. Likewise, the time-like translational symmetry, represented by the existence of Killing fields along the time-like coordinate, is associated with the global conservation of energy. There is global conservation of energy only if there exists an isometry of the metric along the time-like direction.[13] Without these symmetries, the notion of a global conservation law simply loses its meaning. If a spacetime does not have the requisite symmetries, there can be no such conservation laws for that spacetime.

Now we are in a position to consider the metrics of black holes, as well as their symmetries and associated conservation laws. The most common metrics associated with black holes (and the ones used in the proof of Hawking radiation to be discussed later) *do* have some degree of symmetry, allowing us to define global conservation laws on spacetimes describing such black holes. As we will see, these laws are essentially tied to a key property of these black hole metrics: notably, it requires their *time-independence.*

For instance, the Schwarzschild metric describing the vacuum asymptotically flat exterior of a non-rotating uncharged spherically symmetric black hole in Schwarzschild coordinates (t, r, θ, ϕ) is given by:

$$ds^2 = -(1 - \frac{2M}{r})dt^2 + (1 - \frac{2M}{r})^{-1}dr^2 + r^2(d\theta^2 + sin^2\theta d\phi^2) \tag{4.7}$$

where M is the mass parameter and $2M$ is the Schwarzschild radius which determines the event horizon. Importantly, M is constant in time here, i.e. a time-independent parameter. (Schwarzschild 1916, 2) The metric components are independent of t and ϕ, and coordinate transformations reveal two more spatial rotational symmetries. Together the Schwarzschild metric has 4 associated Killing fields – one time-like and three space-like – resulting in global conservation of both energy and angular momentum in the usual spatial directions. Note, however, that the independence of the metric from the time-like coordinate also entails that

[13] See, also, Brown (forthcoming) who details the equal footing of both Noether's theorem and its converse. This suggests a strict two-way relationship between conservation laws and symmetries, though some technical caveats apply.

the metric *never changes in time* – it is *static*.

The Kerr metric, which describes the vacuum asymptotically flat exterior of a *rotating* uncharged *axially* symmetric black hole, is given in Boyer-Lindquist coordinates (t, r, θ, ϕ) by:

$$\begin{aligned} ds^2 = & -\frac{\Delta - a^2 sin^2\theta}{\rho^2}dt^2 - 2a\frac{2Mrsin^2\theta}{\rho^2}dtd\phi \\ & + \frac{\rho^2}{\Delta}dr^2 + \frac{(r^2+a^2)^2 - a^2\Delta sin^2\theta}{\rho^2}sin^2\theta d\phi^2 + \rho^2 d\theta^2 \end{aligned} \tag{4.8}$$

where $\Delta \equiv r^2 - 2Mr + a^2$, $\rho^2 \equiv r^2 + a^2cos^2\theta$, J is the angular momentum parameter and M is again the mass parameter. The event horizons occur where $\Delta = 0$. Once more, by inspection, we can see that the Kerr metric is independent of the time-like t as well as the space-like ϕ. However, since the black hole is rotating, there is a privileged axis of rotation which rules out the two other Killing fields associated with the spherically symmetric Schwarzschild black hole. So the Kerr metric only have two Killing fields, one time-like and one space-like.[14] As a result, we have global conservation laws for energy and angular momentum (in one direction). Since the Kerr metric involves rotation, it is not static unlike the Schwarzschild black hole which describes absolutely no motion. Yet, as with the Schwarzschild metric, the time-independence means that the Kerr metric is likewise *not changing in time* - such a metric is not static, but nevertheless *stationary*.[15] In what follows, for simplicity, I will use 'stationarity' to refer to both staticity and stationarity as discussed here.

4.3 From Hawking radiation to black hole evaporation

With a grasp on what stationarity amounts to in the general relativistic context, let us now examine how the idealization of quasi-stationarity is required for Hawking's 1975 argument from the existence of Hawking radiation to black hole evaporation.

The key idea for Hawking radiation is that we can consider how quantum matter fields

[14] There is also a Killing *tensor* field, though I will not discuss it in this context.

[15] A generalized family of stationary metrics is the Kerr-Newman family of metrics, which also allows one to discuss charged black holes.

behave near a collapsing star as the latter forms a black hole and event horizon, and how this behavior appears to observers during ‘late times’ at infinity, i.e. after the black hole has settled into a *stationary* or *static* state. Hawking (1975) found that the stationary black hole horizon can be interpreted as emitting radiation – and hence having a temperature T – proportional to its surface gravity κ:

$$T = \frac{\kappa}{2\pi} \tag{4.9}$$

This has been derived in a variety of ways.[16] Hawking’s original calculations considered only spherically symmetric collapse, though, as Wald (2001, 12) notes, the effect obtains for any arbitrary gravitational collapse into a black hole. Note, however, that a common assumption remains despite the myriad of generalizations available nowadays: the radiating black hole is assumed *stationary.* (Wald 2001, 12). We’ll return to this in §4 and §5.

Let us assume that Hawking’s derivations establish that black holes can be thought as emitting some amount of energy via Hawking radiation, and that this radiation *can* be interpreted as the temperature of the black hole. Even so, these derivations do not yet amount to an conclusive argument that black holes *really* have a temperature. As Wallace notes,

> these derivations in of themselves do not suffice to establish that Hawking radiation is fully analogous to ordinary thermal radiation, because they imply nothing about whether a radiating black hole ultimately decreases in mass and, thus, surface area. (2018, 11)

Ordinary thermodynamic objects lose energy and/or mass when radiating or losing energy to their surroundings. Even supposing that black holes do radiate energy to their surroundings, do they lose energy or mass as a result? In other words, do they evaporate? To complicate things, Wallace also observes:

> given that there is no robust local definition of gravitational energy and, relatedly, no robust way to understand total energy as a sum of local energies, we cannot simply appeal to a local conservation law to conclude that radiating black holes

[16] See e.g. Hawking (1974 and 1975), Wald (2001) or Carroll (2019).

> evaporate. (2018, 11)

Because there is no robust local notion of gravitational energy,[17] we cannot appeal to the existence of a locally conserved energy to support the argument for black hole evaporation since gravitational energy is precisely what black holes – purely gravitational objects on the classical relativistic understanding – are made of.

Another more pristine way to think about the non-local – *global* – nature of black hole evaporation and Hawking radiation – especially in Hawking's original 1975 derivation – is that his calculation was explicitly *semiclassical*: that is, black holes are classically relativistic objects, against which quantum fields interact. In this context, black hole event horizons are defined simply as "the boundary of the causal past of future null infinity". As Curiel explains:

> The definition thus states in effect that a spacetime has a black hole if one can divide the spacetime into two mutually exclusive, exhaustive regions of the following kinds. The first, the exterior of the black hole, is characterized by the fact that it is causally connected to a region one can think of as being 'infinitely far away' from the interior of the spacetime; anything in that exterior region can, in principle, escape to infinity. The second region, the interior of the black hole, is characterized by the fact that once anything enters it, it must remain there and cannot, not even in principle, escape to infinity, nor even causally interact in any way with anything in the other region. The boundary between these two regions is the event horizon.

As Curiel then puts in plain terms:

> This definition is global in a strong and straightforward sense: The idea that nothing can escape the interior of a black hole once it enters makes implicit reference to all future time—the thing can never escape no matter how long it tries. Thus, in order to know the location of the event horizon in spacetime, one must know the entire structure of the spacetime, from start to finish, so to speak, and all the way out to infinity. As a consequence, no local measurements one can make can ever determine the location of an event horizon. (2019, 29)

Given that Hawking radiation – in Hawking's original argument – essentially depends on global structures such as the event horizon, local notions of energy (and their conservation)

[17] The most prominent proposal is the pseudotensor approach. For a recent proponent of this approach, see Read (2020). For (what I think are successful) rejoinders and partial rejoinders, see Duerr (2019) and De Haro (2022) respectively.

cannot be operative here.

To my knowledge, though, many physicists rely implicitly or explicitly on *some* notion of energy conservation when discussing the back-reaction of Hawking radiation on black holes. A generic answer is that black holes lose energy when Hawking radiation occurs, because Hawking radiation carries energy away to infinity. The assumption here is that energy is conserved somehow, so that any change in energy must be compensated by a corresponding change elsewhere. Hawking (1975) himself phrases the reasoning as follows:

> [Hawking radiation] will give positive energy flux out across the event horizon or, *equivalently*, a negative energy flux in across the event horizon. [...] This negative energy flux corresponding to the outgoing positive energy flux will cause the area of the event horizon to decrease and so the black hole will not, in fact, be in a stationary state. (1975, 219, emphasis mine)

The 'equivalence' here amounts to some reasoning relying on conservation of energy. Hawking (1976) uses this reasoning more explicitly:

> Because this radiation carries away energy, the black holes must presumably lose mass and eventually disappear. (Hawking 1976, 2461)

Wald (2001) says the same:

> Conservation of energy requires that an isolated black hole must lose mass in order to compensate for the energy radiated to infinity by the Hawking process. (2001, 16)

Carlip (2014, 20) simply equates the change of mass of a black hole to the power radiated by the black hole:

$$\frac{dM}{dt} = -\epsilon\sigma T^4 A \tag{4.10}$$

where ϵ is the emissivity parameter, σ is the Stefan-Boltzmann constant, A is the area of the black hole, and T is its temperature. The right-hand side is simply the Stefan-Boltzmann law for the total power radiated by a black body over some surface area, but the claim that this *equates* to the change in mass requires conservation reasoning.

More recently, Carroll states:

> Nevertheless, even in quantum mechanics we have conservation of energy (in the sense, for example, of a conserved ADM mass in an asymptotically flat spacetime). Hence, when Hawking radiation escapes to infinity, we may safely conclude that it will carry energy away from the black hole, which must therefore shrink in mass. (2019, 417)

Note that ADM (Arnowitt-Deser-Misner) mass is conserved (indeed, can only be defined) in asymptotically flat spacetimes because of the asymptotic symmetries of said spacetime.[18] Asymptotically flat spacetimes can be considered to have the symmetries of Minkowski spacetime (and can be described by the flat Minkowski metric) at spatial infinity, and these symmetries allow for the existence of a time-like Killing field. This then gives us a global conservation law for energy, from which we can, again, motivate black hole evaporation.[19]

How are we to interpret all these claims about the use of conservation of energy to infer that black hole evaporation occurs, in light of the inability to employ a local conservation law of energy for such contexts? It seems that the next natural option is to appeal to some *global* conservation law for energy, akin to the one we've discussed above in §2 using Noether's theorem and global symmetries of certain spacetimes. Indeed, this is precisely what Hawking does. He proposes that:

> Although it is probably not meaningful to talk about the local energy-momentum of the created particles [at or near the event horizon], one may still be able to define the total energy flux over a suitably large surface. (1975, 216)

Of course, in this context, the 'suitably large surface' is the event horizon, and so, what Hawking proposes is essentially this: instead of considering the conservation of local notions of energy (e.g. over some worldline/local region of spacetime), we might still be able to talk about *global* notions of energy (and how they are conserved). However, a naïve appeal to the global conservation of energy leads quickly to contradiction.

[18] Roughly speaking, we can define ADM mass as the total deviation of the actual spacetime metric from the flat Minkowski metric. See Arnowitt, Deser and Misner (1962) for more details.

[19] I won't discuss asymptotic flatness in this paper, as that will take us too far afield; the focus here is on the justified use of quasi-stationarity as an idealization. In a follow-up paper, I assess whether asymptotic flatness might face similar issues to the problems raised here for quasi-stationarity.

4.4 A naïve dilemma for black hole evaporation

Since Hawking's argument for black hole evaporation is made in the semiclassical context, the operative assumption is that the classical rules of general relativity hold. This include rules for when a global conservation law exists, i.e. when a spacetime is stationary. (see §2.) From a naïve perspective, there is an underlying tension in the above discussions. Simply put, situations where evaporation is expected to occur are precisely those where there is no global conservation law for energy. On the contrary, situations where we can take there to be a global conservation law for energy are those where evaporation is impossible. So from this naïve perspective, the argument for black hole evaporation cannot take off.

Let me reformulate the worry in the form of a dilemma. For any spacetime in which we want to argue for the occurrence of black hole evaporation, we begin by noting that any such spacetime is either stationary, or not. In other words, the metric for that spacetime is either time-independent or not.

On the one hand, if the spacetime is stationary, then it has the appropriate Killing field structure as we have already seen from §4.2. For example, the Kerr and Schwarzschild metrics are time-independent and are associated with global time-like Killing fields. Yet, we have also seen that stationary metrics are precisely those that do not change over time. Since they do not change over time, the black hole being described by said metric does not change over time either. This entails that black holes described by a stationary metric does *not* evaporate: evaporation entails a change of mass over time (as Carlip (2014) puts it explicitly above), rendering mass a time-*dependent* parameter. Since mass features in the metric, its time-dependence entails the time-dependence of the metric. Black hole evaporation cannot occur for a black hole described by a stationary metric.

On the other hand, if a spacetime is *not* stationary, then parameters of this spacetime – such as mass – are allowed to be time-dependent. They *can* change over time and therefore we can describe the evaporation of a black hole with such a metric. For instance, mass can be a parameter that depends on the time-like coordinate. A simple example discussed by Wallace (2018) is the Vaidya spacetime, where the (retarded time) metric is:

$$ds^2 = -(1 - \frac{2m(u)}{r})du^2 - 2dudr + r^2(d\theta^2 + sin^2\theta d\phi^2) \tag{4.11}$$

which looks a lot like the Schwarzschild metric with a time-dependent mass parameter. However, in these cases we do *not* have a global time-like Killing field for the spacetime in question. The actual metric is *time-dependent*.

Spacetimes which allow for evaporation are precisely spacetimes that do *not* have global conservation of energy due to the lack of a global time-like Killing field. Since there is no global conservation law to rely on, we *cannot* simply assume the black hole experiences a loss of mass or energy as compensation for the positive energy flux due to Hawking radiation. Put another way, spacetimes which *can* accommodate evaporation are precisely those where we have *no justification* for thinking that evaporation occurs.

The naive use of global conservation of energy thus leads us to a dilemma. On both horns of the dilemma, we lack justification for believing in black hole evaporation.

4.4.1 Quasi-stationarity

Hawking (1975) himself was aware of this problem. Evaporation can only occur for a *non-stationary* black hole. However, the black hole was assumed to be *stationary* in the derivation of Hawking radiation. How can we reconcile the two? In response, he argues that we can resolve this problem by employing the idealization of *quasi-stationarity*:

> This negative energy flux will cause the area of the event horizon to decrease and so the black hole will not, in fact, be in a stationary state. However, as long as the mass of the black hole is large compared to the Planck mass 10^5 g, the rate of evolution of the black hole will be very slow compared to the characteristic time for light to cross the Schwarzschild radius. Thus, it is a reasonable approximation to describe the black hole by a *sequence of stationary solutions and to calculate the rate of particle emission in each solution.* (Hawking 1975, 219, emphasis mine)

While we think that actual black holes will *not* be in a stationary state, we might think that they *approximately* undergo a quasi-stationary process. Quasi-stationarity is the idealization that a black hole changing over time can be approximated with a sequence of stationary (or static) solutions. Here's how an explicit rendition of Hawking's defense might look like. We start by accepting that the black hole (and spacetime) in question is actually time-*dependent* and *non-stationary*. However, because its mass is changing so slowly in time when it has a large mass relative to the Planck scale, we can assume that this time-dependent black hole approximates a certain idealization – a sequence of time-independent stationary black holes that can be said to be *quasi-stationary*. For each such stationary metric, there is global conservation of energy associated with that solution. So we can derive Hawking radiation in this stationary regime, and using conservation reasoning here, conclude that the energy of the black hole must decrease.[20] This is where we calculate the aforementioned 'negative energy flux'.

This decrease, of course, cannot happen within *any* of the states in the sequence of stationary black hole solutions which form the quasi-stationary process. Instead, we say that this decrease applies to the *actual* time-dependent black hole for which evaporation *can* happen. We do this by perturbing the mass parameter of the Schwarzschild metric 'by hand', bringing it from one stationary state to another in a temporal sequence of stationary states. This is the quasi-stationary process representing black hole evaporation. In short, we perform the conservation reasoning in the idealized stationary regime but apply the results of this reasoning to the target black hole being modeled.

4.5 Norton's Impossible Process: Quasi-Staticity

As mentioned in the beginning of the paper, this move is reminiscent of one commonly found in equilibrium thermodynamics. There, as is well-known, equilibrium states are those

[20] More concretely, Hawking argues that despite the non-uniqueness of defining local energy-momentum operators (in the so-called 'pseudotensor' approaches to energy in general relativity), we can restrict attention to those operators which both obey local conservation of energy (i.e. eq. 2) *and* are stationary with respect to time-like Killing vectors, i.e. obey global conservation of energy. But time-like Killing vector fields are only well-defined for stationary spacetimes, as we already saw in §4.2.

whose thermodynamic parameters (e.g. volume, pressure, temperature, etc.) do not change with time. That is, equilibrium states are time-independent states. Yet, thermodynamics frequently make use of *quasi-static* processes consisting of sequences of these equilibrium states.[21] For instance, the famous Carnot cycle describing a system interacting with two heat baths with differing temperatures can be described on a pressure-volume diagram, where each point is a state with unchanging pressure and volume. (see Figure 4.1)

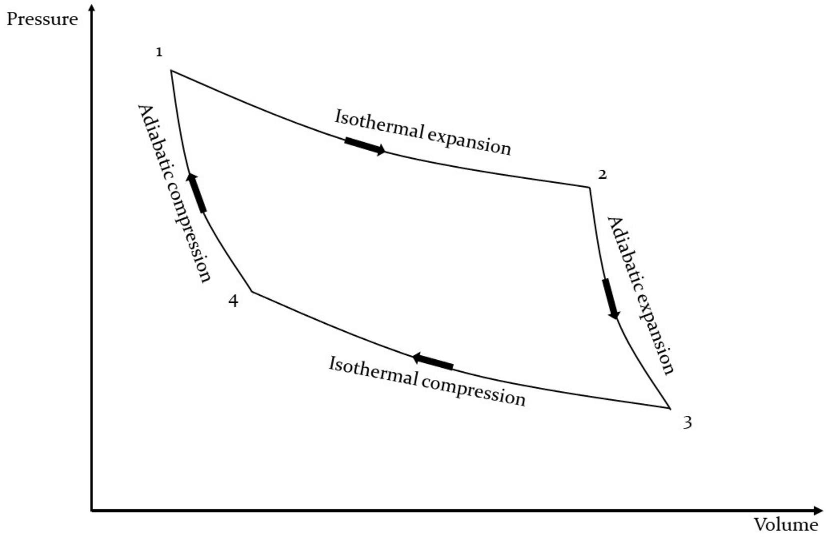

Figure 4.1. The typical Carnot cycle. The points on this graph are equilibrium – time-independent – states.

In these cases, we are interested in systems which are really *time-dependent*. Typical thermodynamic objects change over time: our cup of coffee cools down – and our mug of beer warms up – over time. Nevertheless, if changes to these systems are slow and small enough, they can be modeled by quasi-static processes, such that we can treat them as *effectively time-independent* at any point (or short interval) of time. We can then perform all thermodynamic calculations in terms of this quasi-static idealization, while keeping in mind that these calculations should really apply to the actual time-dependent system. This appears to vindicate Hawking's argument for black hole evaporation – we assume quasi-stationarity

[21] See, for instance, Carathéodory (1909, 366).

for conservation reasoning, and then apply the fruits of this reasoning to the actual system, which is just what we already do in classical thermodynamics.

Unfortunately, quasi-static processes in classical thermodynamics are not conceptually innocent. By inspecting why they work for classical thermodynamics, we can see why the analogous use of quasi-stationarity in the case of Hawking's argument for black hole evaporation does not.

As Norton (2016) recently argued, quasi-static processes, too, come with their own internal tensions. Quasi-static processes are constituted by sequences of equilibrium states, each of which is approximated by an actual physical system at some time. These sequences are meant to be sequences of states *in time* – curves on thermodynamic state space (e.g. Figure 1) are parametrized by a time-like parameter. Furthermore, typical quasi-static processes describe variations in equilibrium states. Figure 1, for instance, includes curves with varying volume and pressure, although each point on these curves represents an equilibrium state. In other words, we are supposed to envision changes to these equilibrium states over time. Yet, equilibrium states do not change with time *by definition*. So on the face of it, quasi-static processes change over time, but also do not change over time – an outright contradiction.

The correct way to interpret quasi-static processes, of course, is not to take them to be *actual* processes with exactly those properties discussed above. Rather, Norton shows how actual systems can undergo processes which are approximately (but never literally) quasi-static processes. The distinction between approximation and idealization is a subtle one, though one distinction was recently introduced by Norton (2012). An approximation is characterized by being "an inexact description of a target system" which is , while an idealization is "a real or fictitious system, distinct from the target system, some of whose properties provide an inexact description of some aspects of the target system." (Norton 2012, 209) Norton gives one simple example of this distinction: that of a body of unit mass falling in a weakly resisting medium. Its velocity v at time t is:

$$\frac{dv}{dt} = g - kv \tag{4.12}$$

where g is the acceleration due to gravity and k is a coefficient representing friction. When falling from rest at $v = t = 0$, its velocity can be expanded as:

$$v(t) = \frac{g}{k}(1 - e^{-kt}) = gt - \frac{gkt^2}{2} - \frac{gk^2t^3}{6} - ... \tag{4.13}$$

When there is low friction (i.e. k is small), the fall of the ball is almost exactly described by:

$$v(t) = gt \tag{4.14}$$

In terms of Norton's distinction, we can say that (14) inexactly describes the ball's descent. However, we can, in Norton's terms, 'promote' this approximation to an idealization by having (14) directly refer to a fictitious system, that of a ball falling in a vacuum such that $k = 0$. Hence, (14) exactly describes such a system though the system need not exist (i.e. (14) is here an idealization), while it only provides an inexact description when the system is not in vacuum (i.e. (14) is here an approximation).

Using this distinction, we may understand quasi-static processes, taken literally, as idealizations – they can only be fictitious systems since they are contradictory in nature. They may be approximated by actual systems, of course, but since we are typically more interested in the dynamical properties of these systems, the processes being approximated by quasi-static processes will be time-dependent ones. Sets of these time-dependent processes may come arbitrarily close to quasi-static processes by having vanishingly small driving forces – differences in temperature, pressure or other thermodynamic parameters – but no actual process ever has all the exact properties of quasi-static processes. We cannot simply 'take the limit' and let the driving forces actually go to zero, for we have already seen how that leads to contradiction. As Norton puts it:

> We have a sequence of [non-equilibrium] processes, each of which is slowed by diminishing the driving forces. Each process carries the property of completing a change, while requiring ever more time to do it. The limit of this property is the property of completing a change. The limit approached by the processes themselves, however, is no process at all. It is merely a static set of states in equilibrium that no longer carry the limit property of completing the change. (2016, 46)

Why, then, is the use of quasi-staticity still so widespread, despite the issues we have discussed? This is because quasi-static processes *do* approximate actual processes, even though actual processes are *never* exactly quasi-static. For large systems, the time-dependent driving forces are so minuscule relative to the dynamics of the system that we may neglect them for all practical purposes and model them (inexactly!) with quasi-static processes, though we must remember that these systems are *not* undergoing quasi-static processes.

Real systems are simply undergoing time-dependent processes which are approximated by quasi-static processes in an inexact fashion. Importantly, Norton shows how we may recover key thermodynamic results about the efficiency of reversible heat engines, absolute temperature, and the Clausius inequality for thermodynamics, by *working purely with time-dependent processes*, without getting led astray in pathological cases like processes at the molecular scale.[22] All this is to say that the exact properties of quasi-staticity – of a system literally experiencing change and no-change – are never *essential* to doing thermodynamics. This is good news, since a system bearing the exact properties of quasi-staticity never exists, being contradictory in nature.

4.5.1 Idealization & de-idealization

The key lesson of the foregoing is simple but important: we must not confuse properties of the idealization with properties of the system approximating said idealization. Quasi-static processes contains properties of change and no-change, but only at the limit of zero driving force. Thankfully, it turns out that actual thermodynamic systems never have exactly those

[22] Due to space constraints, I will not rehearse Norton's derivations here. See Norton (2016, §4) for detailed analysis.

properties, and thermodynamics does not *essentially* require those properties of quasi-staticity. That's why we may continue using quasi-static processes *qua* approximation.

The broad spirit of this lesson is neither new nor unique to Norton's works. As McMullin (1985) already wrote, idealizations are typically justified by a clear *de-idealization* process demonstrating how we may remove the simplifications and distortions introduced in the idealization in order to describe realistic systems. For instance, consider Bohr's model of the hydrogen atom: an idealization assuming a stationary and infinitely massive proton, with an electron orbiting in a perfect circle – both assumptions which were known or suspected to be false. As McMullin discusses, a de-idealization in this case might require the provision of a more realistic model in which the proton could undergo motion, or the electron followed a more general elliptical orbit. If we can show how this more realistic model might approximate the idealized one, we may then justifiably continue to use the idealized Bohr model despite its false assumptions provided we know when those approximations are good (relative to some standard).

A kindred idea is explored by Fletcher (2020a) who discusses Duhem's *principle of stability* as an plausible epistemological principle in scientific modeling. Roughly, the principle states that we are justified in inferring, from a model of some phenomenon, conclusions about said phenomenon only if these conclusions remain approximately true of the actual phenomenon when the modelling assumptions only approximately hold.[23]

This assumes, of course, that there is *some* way of describing *how* these modelling assumptions approximately hold. In turn, the principle of stability dovetails with McMullin's discussions of de-idealization: if we can show that there exists some de-idealization procedure for a model in question, we can show how this model approximates some target system despite its idealized and false assumptions. This then takes us some ways towards a justification for using the model to make inferences about the world. Conversely, though, if we *cannot* show how the model approximates a target system and relevant results from the model about the

[23] For more detailed discussion, see Fletcher (2020a).

target system essentially turn on the model's idealized assumptions, then we are *not* justified in making conclusions about real-world phenomena from said model.

In the case discussed above, quasi-static processes are idealizations but they can also be adequately de-idealized: Norton tells us how they approximate actual non-quasi-static processes with non-vanishing driving forces. Hence, while quasi-static processes *qua* idealizations are useful tools, they are not essential to the description of realistic systems – as Norton shows, there is always a de-idealization procedure available in principle.

This consideration of whether a property of an idealized system is *essential* to the description of actual systems approximating said idealization has already been much discussed in the literature on the philosophy of thermodynamics. Concerning the physics of phase transitions, Callender (2001), Butterfield (2011), Menon and Callender (2013) and Palacios (2019) have all argued that the idealization of a system exactly at the thermodynamic limit is not essential to the understanding of phase transitions, and so the phenomenon of phase transitions does *not* require the actual existence of an idealized system (in this case, a system at the thermodynamic limit with infinite number of particles). Again, a de-idealization procedure is available. This is good news, because real systems undergoing phase transitions are finite and never have the exact (infinite) properties of the idealization. Likewise, as we have seen for the case of quasi-staticity and thermodynamics, quasi-staticity is not essential to thermodynamics. Actual processes are never quasi-static, though they might approximate certain features of quasi-static processes.

4.6 Quasi-stationarity and (the lack of) de-idealization

So how does the above discussion bear on quasi-stationarity? To start, the same issue with quasi-static processes arises here for quasi-stationary processes: real black holes *cannot* literally be undergoing quasi-stationary processes. Quasi-stationary processes do not refer to actual processes since black holes (or any process) cannot be both time-independent and time-dependent. (See §4.4.) As such, quasi-stationarity is an idealization. At best, a system may be inexactly described – *approximated* – by quasi-stationary processes.

We may then ask: is there a de-idealization procedure by which black holes are approximated by quasi-stationary processes, similar to what Norton has done? Furthermore, can this procedure give us something like the global conservation of energy without essentially requiring the limit property of quasi-stationarity – of change and no-change at once? This will depend on what the idealization is being used for, so that we can know what properties of the idealization are being employed.

In the remainder of this section, I argue that Hawking's 1975 use of the quasi-stationary idealization remains unjustified, because in this context there is no clear de-idealization procedure for the specific property of quasi-stationary processes being employed by Hawking's argument: the global conservation of energy. The idealization of quasi-stationarity, just like the idealization of quasi-staticity, cannot always be assumed to be applicable but requires an appropriate justification via de-idealization. In this context, the appropriate de-idealization procedure will tell us how a real system approximates the property of possessing a globally conserved energy. This will require us to show how a realistic spacetime approximates the structure of global time-like Killing fields, with which one may justify the claim that energy is 'approximately globally conserved'. If we can do so, then the argument for black hole evaporation from quasi-stationarity can begin to take off without essentially depending on the idealization of quasi-stationarity.

It seems to me, however, that there is no clear de-idealization procedure for the property of global conservation of energy in quasi-stationary processes. Specifically, I will argue that there is a conceptual difficulty in understanding just in what sense a spacetime approximately has a time-like Killing field (or Killing field in general), in stark contrast to the case of vanishing driving forces approximating quasi-staticity. Yet this seems to be exactly what we need in order to discuss approximate conservation, given the relationship between Killing fields and conservation laws I've sketched in §2.

To start off, there is a general worry against the *very idea* of approximate Killing fields, simply because of the nature of general relativity: there is no fixed maximally symmetric background spacetime structure against which we have a canonical way of considering deviations from symmetry, i.e. 'how close' a spacetime is to being symmetric.[24] For example, there is a natural way in which one can consider deviation from symmetry in Newtonian spacetime (e.g. in terms of asphericity and almost rigid fields). Since Newtonian spacetime is maximally symmetric, flat, and non-dynamical, it provides a fixed background – akin to a spacetime ruler – against which we may measure deviations and closeness to symmetry. However, there is generally no such fixed background in general relativity with which we can construct such a canonical metric of closeness. *Prima facie*, this should warn us against hoping for too much when it comes to seeking approximate Killing fields.

Building on this foundational conceptual worry, and related to this lack of a 'spacetime ruler', are three prominent problems related to various attempts at constructing approximate Killing fields.[25]

Problem 1: 'closest' does not mean 'close'. These procedures typically provide no clear way to understand how 'close' a given vector field is to a Killing field, only that said vector field is, in fact, the closest. This means there is no clear way to understand the deviation from symmetry, and hence evaluate the accuracy of any claims about approximate conservation we might want to make. The fact that the closest In-N-Out burger place to me is in California (as I write from Southeast Asia) is no comfort, for there is no reasonable scale in which it is, in fact, close to me. The situation is worse here, since there isn't even a way to appropriately characterize the 'closeness' of a given vector field to a Killing field, unlike a ruler (or its ilk) for the distance between me and the nearest In-N-Out. Without such a notion of 'distance',

[24] For a cognate discussion, see Fletcher (2014).

[25] Fletcher (2020b) has recently provided an account of how spacetimes might *locally* approximately possess certain spacetime symmetries, though Fletcher acknowledges that this account does not extend yet to the global case. Since we are concerned here precisely with what an approximately *global* symmetry and its associated conserved quantity – time-independence and a globally conserved energy – would look like, that account does not suffice quite yet. But I leave open – and welcome – the possibility of future developments.

it's not clear to me how one can understand the 'approximation' relation, since discussions of approximate symmetry typically employ some such distance function or at least some similarity relation.[26]

Many extant procedures seek to find the 'next best thing' to Killing fields by some generalization of Killing's equation, which, if we recall, is:

$$\nabla_\nu \xi_\mu + \nabla_\mu \xi_\nu = 0 \tag{4.15}$$

As mentioned in §4.2, a Killing field satisfies this equation. However, for spacetimes without Killing fields, the equation generally has no nontrivial solutions. (Matzner 1968, 1657) Instead, these procedures try to find generalized equations to which a Killing field is but one of many solutions. This is supposed to justify the other solutions as suitable generalizations of Killing fields, insofar as they belong to the same class of solutions. For instance, Beetle & Wilder (2014) employs an Euler-Lagrange equation of the form

$$\Delta_K u^\beta = \kappa_u u^\beta \tag{4.16}$$

where what they term the "Killing Laplacian" operator Δ_K is defined as:

$$\Delta_K u^\beta := -2\delta_\gamma^{(\beta} g^{\lambda)\nu} \nabla_\lambda \nabla_\nu u^\gamma \tag{4.17}$$

Here, u^β are vector field solutions to the equation with corresponding eigenvalues κ_u. As they note, Matzner (1968) employs a similar method (though in a slightly different form). This procedure simply defines the 'most' approximate Killing field for any given metric to be the vector field solution with the smallest κ_u greater than zero. (The solution with a vanishing eigenvalue corresponds to an actual Killing field.)

However, the problem is that this procedure fails to provide physical meaning to 'how

[26] See Rosen (2008) for discussion of this distance function understood as a pseudometric, and Fletcher (2021) for more recent discussion.

close' these generalized fields are to Killing fields. As Matzner (1968, 1657) notes, "we do not have to assume the deviation from symmetry is small" when we are looking for the vector field which best approximates a Killing field. In other words, the most approximate vector field need *not* be close to being a Killing field at all. The approach simply finds a discrete spectrum of eigenvalues (and associated vectors), each increasingly 'further' away from being a Killing field. By stipulation, we pick the lowest non-zero eigenvalue and its associated vector field as the most approximate Killing field. Yet, it is not exactly clear in what sense these vector fields are 'close' to Killing fields, beyond the fact that these fields become Killing fields when their associated eigenvalues vanish. Along what dimension are these fields becoming 'closer' to being Killing fields, and how are we to understand this sort of 'distance'? In approximating quasi-static processes with systems possessing small driving forces, we can see how the size of the driving forces dictate the deviation from quasi-staticity. What does the ordering of these generalized vector fields mean here?

As Chua & Callender observe (in the context of deriving time from no-time in quantum gravity): "at the level of pure math, one can 'derive' virtually any equation from any other if one is allowed to assume anything. It makes no sense to say that one equation or quantity is "close" to another absent a metric." (2021, 1176) I think the same is precisely going on here. The approach in question does not provide a physical interpretation of this ordinal ordering of eigenvalues and associated vector fields. The only anchor is the formal fact that the Killing vector field appears as a solution of this more general class of equations, but what does this generalization amount to? Without a convincing story, it remains unclear what it means for a time-dependent physical system, like a black hole, to approximate a spacetime with a Killing field. All we know is that this field is *not* Killing, and that we can define a vector field on it which 'best' captures the properties of a Killing field. But is 'best' enough? Even if it does generate some vector field that is 'close' to Killing, it is still not clear how this is supposed to approximate a Killing field because we are not given a clear understanding of 'closeness'.

A different approach by Cook & Whiting (2007), adopted in some form by Lovelace et al (2008), has the same problem. This approach compares general vector fields to Killing vector fields on 2-spheres and identifies

$$S_{ij}S^{ij} = (\nabla_\mu \nabla_\nu v)(\nabla^\mu \nabla^\nu v) - \frac{1}{2}(\nabla^\alpha \nabla_\alpha v)^2 \tag{4.18}$$

as a term which vanishes if said vector field satisfies the Killing equation (4.15).[27] They then represent how 'far' a vector field is from a Killing field by understanding $S_{ij}S^{ij}$ as an 'error' term and attempting to extremize $S_{ij}S^{ij}$. However, again, their approach only guarantees that $S_{ij}S^{ij}$ is "as close to zero as possible" (Cook & Whiting 2007, 2), but does not provide a clear sense of how close it actually is to a Killing field, and what the meaning of closeness in terms of S_ij amounts to physically. They admit that the usefulness of their results for an approximate Killing vector depends on the extent to which a physically meaningful story can be provided for them "since a Killing vector cannot be produced where one does not exist." (2007, 4) However, they crucially do not provide this story themselves. Yet, a physically meaningful sense in which the S_{ij} term represents 'approximation to a Killing field' seems to be exactly what we need here in the present discussion.

Another popular approach discussed by Bona et al (2005) faces the same problem. There is no clear metric for assessing how 'close' a given vector field is to being a Killing vector field. Bona et al employs yet another generalization of Killing vector fields via what they call the 'almost-Killing' equation. This equation generalizes from the Killing equation by showing that solutions to the Killing equation are also solutions to the almost-Killing equation. Bona et al derives a "wave equation" of the following form:

$$\Box \xi_\mu + R_{\mu\nu}\xi^\nu + (1-\lambda)\nabla_\mu(\nabla \cdot \xi) = 0 \tag{4.19}$$

where $\Box$ denotes the d'Alembertian operator, $R_{\mu\nu}$ is a Ricci term, and λ is a free parameter.

[27] Here, v is a scalar field constructed by Cook & Whiting from the decomposition of a general vector field on 2-spheres.

They then show that, if ξ^{α} is a Killing vector, then it satisfies (4.19). This provides motivation to take (4.19) as the generalization of Killing's equation. However, essentially the same problem as the earlier approaches arises here. As Feng et al (2019, 5) point out, vector fields which satisfies the almost-Killing equation need not be in any meaningful sense 'close' to being Killing vector fields. Notably, Feng et al observes that any transverse-traceless tensor $Q_{\mu\nu}$ of the following form

$$Q_{\mu\nu} := \frac{1}{2}(\nabla_{\mu}\xi_{\nu} + \nabla_{\nu}\xi_{\mu}) \tag{4.20}$$

satisfies the almost-Killing equation. Of course, when $Q_{\mu\nu} = 0$, that is equivalent to saying that the vector field ξ_{ν} satisfies the Killing equation since $Q_{\mu\nu} = 0$ is equivalent to the Killing equation (sans a factor of $\frac{1}{2}$). However, as they point out, $Q_{\mu\nu}$ need *not* vanish, and indeed "the components of $[Q_{\mu\nu}]$ need not be small." (Feng et al 2019, 5) Driving home the point I have been making so far, Feng et al go on remark that "the term 'almost Killing' is somewhat of a misnomer" because it's not guaranteed to be almost Killing at all. As with the other approaches I have examined so far, this approach, too, fails to provide a clear sense in which the solutions to these 'generalized' Killing fields actually approximate Killing fields.

In short, the first problem is that these approaches fail to provide a physically compelling story for what it means for this myriad of proposed vector fields to approximate a Killing field, and hence how one might go about de-idealizing the notion of a globally conserved energy. Given the lack of a 'spacetime ruler' with which to provide a canonical measure of 'distance from symmetry', such a problem might be unsurprising: if there's no canonical measure of 'almost-symmetric' in general, then it might be moot to hope for a clear meaning to the claim that something is 'almost-Killing'.

Problem 2: no guarantee of time-like Killing fields. Many of these procedures for obtaining approximate Killing fields do not guarantee that we can obtain an approximately *time-like* Killing field, only that we can obtain *some* approximately Killing fields.[28] This means these procedures typically do not tell us how close a spacetime is to having a *time-like* Killing field, but only how close a spacetime is to having *some* Killing field at all. Note, then, that on these procedures a spacetime may not turn out to have anything approximating a time-like Killing vector field at all! These procedures do not necessarily help us de-idealize away from a spacetime with a *time-like* Killing field to a spacetime with some approximate time-like Killing field, but only to a spacetime with *some* Killing field. This, however, is of no help when our goal is to find a de-idealization procedure for the energy conservation of a realistic black hole, in terms of an approximate *time-like* Killing field. What we seek here is a story for how any spacetime approximates one with energy conservation, not just one with any conserved quantity.

Problem 3: overly stringent assumptions. Many approximation procedures tend to feature strong assumptions which may not be realistic. As mentioned, one can derive almost anything from anything else if we assumed strong enough assumptions, but "approximations require physical justification." (Chua & Callender 2021, 1176) For instance, Matzner's (and Beetle & Wilder's) procedure requires the assumption that spacetime be compact. This is a *very* strong assumption. For one, an arbitrary spacetime can very well be unbounded and infinite, instead of compact, and it is in fact unclear whether our universe is in fact one or the other. Furthermore, a result due to Geroch (1967) suggests that any compact spacetime admits closed timelike curves.[29] It seems odd that one may only approximate Killing fields in compact spacetimes and not otherwise, given that we may very well live in non-compact spacetimes. As such, these procedures may be lacking in realism, and may not even be applicable to the actual world, in terms of which de-idealization would take place.

Another assumption that frequently shows up is some variant of an assumption of asymptotic flatness. For instance, Matzner's procedure can do without the compactness assump-

[28] See, for instance, Matzner (1968), Cook & Whiting (2007), Lovelace et al (2008) and Beetle & Wilder (2014).

[29] See e.g. Manchak (2013) for discussion of compactness.

tion, if one demands that certain terms of the class of vector fields in consideration vanish "at infinity", which amounts to some assumption of asymptotic flatness. (Matzner 1968, 1658) It also shows up as an assumption in the almost-Killing equation approach as one way to avoid some scathing problems with the approach. As Feng et al (2019) shows, a Hamiltonian analysis of the almost-Killing equation reveals that the Hamiltonian for this equation is generally unbounded from below. As they explain, "an unbounded Hamiltonian generally signals the presence of runaway instabilities, which can potentially drive solutions far from the Killing condition." (2019, 2) This means that the almost-Killing equation approach may generate approximate Killing fields which are not close to Killing at all. This returns us to the first problem.

Feng et al argues that we can avoid this problem for the almost-Killing approach if we assume a vacuum spacetime (for which we can motivate a specific choice of initial conditions, dynamical conditions, and λ parameter, hence obtaining a positive-definite Hamiltonian), or if we impose asymptotic flatness as a condition on the spacetime in consideration. (Feng et al 2019, 8) The former does not model any realistic spacetime and does not help us in our de-idealizing. What of the latter? Contrary to the earlier-discussed assumption of compactness, asymptotic flatness is not quite as controversial given its use in much of black hole physics. As such, it seems like a fairly tame assumption. However, I think that any procedure that *requires* the property of asymptotic flatness only passes the buck: we must then provide an account of whether we can de-idealize *that*.[30]

To sum up, the extant procedures I have examined face three problems. Firstly, and most problematically, there is generally no sense in which a proposed approximate vector field is 'close' to a Killing field. Secondly, even *modulo* the first problem, we are not always guaranteed a time-like Killing field. Finally, some of these procedures involve strong and possibly unrealistic assumptions. This complicates the search for a de-idealization procedure away from quasi-stationary spacetime, by preventing us from stating just *how* we are supposed

[30] For reasons to do with the above, I think that de-idealizing asymptotic flatness will be equally challenging. Questions for approximate Killing fields become questions for approximate *asymptotic* Killing fields. However, as mentioned in the beginning, I will not engage with asymptotic flatness here – I leave it for future work.

to de-idealize the properties of a Killing field, found only in stationary spacetimes, for non-stationary spacetimes. Granted, I have not proven a negative existential claim here: there *could* very well be a satisfactory procedure in the future. Nevertheless, the above concerns provide strong reasons to be concerned that Hawking's argument for black hole evaporation might be lacking justification, insofar as we cannot explicate what it means for there to be an approximately globally conserved energy via approximate Killing fields.

Taken together with the general worry that there is not a canonical measure of 'deviation from symmetry' above, these problems should pose a significant challenge to anyone looking for an adequate de-idealization procedure by appealing to approximate global conservation of energy via approximate Killing fields.

Barring such a de-idealization procedure, it seems that Hawking's move (from §4.4.1), which seeks to employ the global conservation of energy in a quasi-stationary process to argue for black hole evaporation, *essentially* needs the idealization of quasi-stationarity, in which a time-dependent system has time-independent properties, viz. that of having a conserved energy. We don't yet have a clear sense of what it means for realistic spacetimes to approximate such a property. From the perspective of Duhem's principle of stability, we lack justification in employing such arguments when discussing realistic systems in our actual world, since we have yet to find a way to understand what it means for a black hole to be approximated by a quasi-stationary process such that vector fields in a non-idealized black hole spacetime approximates Killing fields. Hence, we cannot yet make the argument for black hole evaporation without relying on the existence of a time-like Killing field and associated global conservation of energy. However, since this argument for black hole evaporation essentially requires both time-dependence and time-independence, it can only be described by a process that has both properties. In other words, it can only be described by an idealized quasi-stationary process, but such processes do not exist on pains of contradiction.

As such, Hawking's argument for black hole evaporation from quasi-stationarity cannot take off yet *sans* an appropriate de-idealization procedure.

4.7 Quasi-Stationarity for the Sun

So far, I have argued that quasi-stationarity isn't always justified: at least for Hawking's argument, its use is unjustified because the argument depends essentially on a specific limit property of quasi-stationary processes for which no de-idealization is available: global conservation of energy for a dynamical black hole.

However, this is not to say that quasi-stationarity is *never* justified. Consider one case: using the quasi-stationary idealization for modeling the Sun despite its dynamical nature. This is sometimes called the *standard solar model* (SSM), and is generally held to be one of our best references in astrophysics. (Turck-Chièze 2016) On this picture, "the star is spherical, described by a succession of hydrostatic equilibria, and without effects of rotation and magnetic field." (Turck-Chièze & Couvidat 2011, 8) In other words, the Sun (or stars in general) is idealized as undergoing a quasi-stationary process – going through sequences of stationary states. The ingredients in this model include the following:

1. The Sun is assumed to be in hydrostatic equilibrium: the 'outward' radiative and particle pressures exactly counteracts the 'inward' pull of gravity.

2. Energy transport in and on the Sun is solely by photons or convective motions (and hence ignoring gravitational effects).

3. Energy generation in the Sun is solely via nuclear reactions.

4. The abundance of elements in the interior of the Sun is due solely to nuclear reactions (ignoring, again, gravitational effects because "they are estimated to be small over the lifetime of the Sun").

This model can then be used to predict neutrino fluxes from the Sun, the abundance of certain elements on it, or estimate the age of the Sun by fitting the model with the current parameters of the Sun (e.g. luminosity, radius).

One might worry that the issues I have raised about quasi-stationarity – that it is a literally impossible process – for Hawking's argument for black hole evaporation might 'infect' other parts of astrophysics. After all, we know that the Sun is *not* stationary or actually in equilibrium. So are we likewise unjustified in modelling the Sun with quasi-stationary processes via the SSM, e.g. to assume the conservation of energy on the Sun? This would amount to a reductio ad absurdum of my argument, since we *do* think we are justified in modelling the Sun with quasi-stationary processes such as in the SSM.

However, I think my worries can be suitably quarantined to Hawking's argument for black hole evaporation and, hence, do not render standard usage of this quasi-stationary idealization, e.g. for the Sun via the SSM, unjustified.

The crucial difference is this: in the case of the SSM and the Sun, typical arguments using the model do *not* essentially depend on global conservation of energy. When the conservation of energy is used in the model for making predictions about neutrino fluxes or the abundance of certain elements, it is in *local* processes *par excellence*: process concerning nuclear reactions or energy transport by photons or convective motion. Indeed, these are precisely the sort of processes which are "determinable by local observations of phenomena of the sort that are the bread and butter of astrophysics." (Curiel 2019, 28) They do not appeal to global space-time structures such as the event horizon. But conservation of energy for local processes *always* holds in general relativity, even away from highly symmetric models of spacetime such as stationary solutions. De-idealization away from the SSM, by treating the Sun as dynamical and out of (but close to) equilibrium, adding in magnetic fields or rotation,[31] for example, does not require us to provide a de-idealization of the global conservation of energy, since standard uses of these models do not require global conservation of energy. In other words, the de-idealization of the SSM, for its standard uses, does *not* run into the problem with approximate Killing fields contrary to the case of Hawking's argument for black hole evaporation. The limit property of

[31] See Turck-Chièze & Couvidat 2011 and their 'seismic solar model' for a concrete examples of how such de-idealizations might work out.

quasi-stationary processes – global conservation of energy for a dynamical system – is not required nor essential to discussing the Sun, since the SSM can be de-idealized without appeal to global conservation of energy.

On the contrary, we've seen how Hawking's argument for black hole evaporation essentially relies on the global conservation of energy, since calculations about Hawking radiation in the black hole context must rely on global structures such as the event horizon for which there is no feasible local notion (and hence no local conservation law), as noted in §4.3. The crucial – if obvious – point here is that the very same idealization can be justified in one case and not in another, depending on the properties of the idealization we are using to make inferences about the system being modeled.

4.8 Conclusion

Summing up, I have argued that a widely used idealization for black hole physics – quasi-stationarity – shares much in common with quasi-static processes: both are literally 'impossible processes'. To my knowledge, no one has made this connection between the two idealizations. Crucially, this means that the justified use of these idealizations cannot be taken for granted. A story for why we are allowed to use such idealizations to model realistic systems must be provided, especially in the case of black hole physics and Hawking radiation where we do not even have empirical evidence against which our models can be calibrated or evaluated. In most other cases, this story can be available in principle, but in the case of Hawking's argument for black hole evaporation, no clear procedure is at hand yet for de-idealizing a property of quasi-stationary processes required for the argument – a globally conserved energy – and so the argument is not yet justified.

Importantly, this does not mean that the idealization of quasi-stationarity can never be used in a justified manner. As I have attempted to show for the case of the Sun and the SSM, the use of quasi-stationarity might still be justified if such usage does not rely on limit properties of the quasi-stationary idealization such as global conservation of energy for a dynamical system. The main uses of the SSM depend on local processes within the Sun, and does not depend essentially on global conservation of energy. Hence, my criticisms of Hawking's argument does not infect other uses of quasi-stationarity *per se*, and can be quarantined to uses of this idealization which depend essentially on its limit properties.

More generally, the foregoing emphasizes the need to be careful in employing idealizations in physics and elsewhere; the use of any idealization in making inferences about the world must be 'checked' depending on what we want to use it for, and what properties of the idealization are being utilized. At the very least, we should have a story for how this idealization can approximate reality in some concrete sense – a de-idealization procedure. This runs counter to recent proclamations by philosophers who argue that idealizations are and should go 'unchecked': the risk of not 'checking' one's use of idealizations is that one risks being lost in the idealized model with no anchor to reality.[32]

[32] See e.g. Potochnik (2017).

Bibliography

Abramowicz, Marek A. and P. Chris Fragile (Jan. 2013). "Foundations of Black Hole Accretion Disk Theory". In: *Living Reviews in Relativity* 16.1, p. 1. ISSN: 1433-8351. DOI: 10.12942/lrr-2013-1. URL: https://doi.org/10.12942/lrr-2013-1.

Arnowitt, Richard L., Stanley Deser, and Charles W. Misner (1962). "The Dynamics of general relativity". In: *Gen. Rel. Grav.* 40, pp. 1997–2027. DOI: 10.1007/s10714-008-0661-1. arXiv: gr-qc/0405109.

Beetle, Christopher and Shawn Wilder (Mar. 2014). "Perturbative stability of the approximate Killing field eigenvalue problem". In: *Classical and Quantum Gravity* 31.7, p. 075009. DOI: 10.1088/0264-9381/31/7/075009. URL: https://doi.org/10.1088/0264-9381/31/7/075009.

Bona, C., J. Carot, and C. Palenzuela-Luque (Dec. 2005). "Almost-stationary motions and gauge conditions in general relativity". In: *Phys. Rev. D* 72 (12), p. 124010. DOI: 10.1103/PhysRevD.72.124010. URL: https://link.aps.org/doi/10.1103/PhysRevD.72.124010.

Brown, Harvey (forthcoming). "Do symmetries "explain" conservation laws? The modern converse Noether theorem vs pragmatism." In: *The Philosophy and Physics of Noether's Theorems.* Ed. by James Read, Bryan Roberts, and Nicholas Teh. Cambridge, UK: Cambridge University Press.

Butterfield, Jeremy (2011). "Less is Different: Emergence and Reduction Reconciled". In: *Foundations of Physics* 41, pp. 1065–1135.

Carathéodory, C. (1909). "Untersuchungen über die Grundlagen der Thermodynamik". In: *Mathematische Annalen* 67.3, pp. 355–386.

Carlip, S. (2014). “Black hole thermodynamics”. In: *International Journal of Modern Physics D* 23.11. DOI: 10.1142/S0218271814300237.

Carroll, Sean M. (2019). *Spacetime and Geometry: An Introduction to General Relativity.* Cambridge University Press. DOI: 10.1017/9781108770385.

Chua, Eugene Y. S. and Craig Callender (2021). “No time for time from no-time”. In: *Philosophy of Science* 88.5, pp. 1172–1184. DOI: 10.1086/714870.

Cook, Gregory B. and Bernard F. Whiting (Aug. 2007). “Approximate Killing vectors on S^2”. In: *Phys. Rev. D* 76 (4), p. 041501. DOI: 10.1103/PhysRevD.76.041501. URL: https://link.aps.org/doi/10.1103/PhysRevD.76.041501.

Curiel, Erik (2019). “The many definitions of a black hole.” In: *Nature Astronomy* 3, pp. 27–34.

De Haro, Sebastian (forthcoming). “Noether’s Theorems and Energy in General Relativity”. In: *The Philosophy and Physics of Noether’s Theorems.* Ed. by James Read, Bryan Roberts, and Nicholas Teh. Cambridge, UK: Cambridge University Press.

Duerr, Patrick M. (Dec. 2019). “Against ‘functional gravitational energy’: a critical note on functionalism, selective realism, and geometric objects and gravitational energy”. In: *Synthese.* ISSN: 1573-0964. DOI: 10.1007/s11229-019-02503-3. URL: https://doi.org/10.1007/s11229-019-02503-3.

Feng, Justin C., Edgar Gasperín, and Jarrod L. Williams (Dec. 2019). “Almost-Killing equation: Stability, hyperbolicity, and black hole Gauss law”. In: *Phys. Rev. D* 100 (12), p. 124034. DOI: 10.1103/PhysRevD.100.124034. URL: https://link.aps.org/doi/10.1103/PhysRevD.100.124034.

Fletcher, Samuel C. (June 2014). *Similarity, Topology, and Physical Significance in Relativity Theory.* URL: http://philsci-archive.pitt.edu/17235/.

— (Nov. 2018). “Global spacetime similarity”. In: *Journal of Mathematical Physics* 59.11, p. 112501. DOI: 10.1063/1.5052354. URL: https://doi.org/10.1063%2F1.5052354.

— (Mar. 2020b). *Approximate Local Poincaré Spacetime Symmetry in General Relativity.* Forthcoming in Claus Beisbart, Tilman Sauer, and Christian Wüthrich, eds. ”Thinking about Space and Time." Einstein Studies, vol. 15, Birkhäuser. URL: http://philsci-archive.pitt.edu/17229/.

Fletcher, Samuel C. (2020a). "The principle of stability." In: *Philosophers' Imprint* 20.3, pp. 199–220.

Geroch, Robert (1967). "Topology in General Relativity." In: *Journal of Mathematical Physics* 8, pp. 782–786.

Hawking, S. W. (1974). "Black hole explosions?" In: *Nature* 248, pp. 30–31.

— (1975). "Particle creation by black holes." In: *Commun. Math. Phys.* 43, pp. 199–220.

— (1976). "Breakdown of predictability in gravitational collapse". In: *Phys. Rev. D* 14 (10), pp. 2460–2473.

— (1977). "The Quantum Mechanics of Black Holes". In: *Scientific American* 236.1, pp. 34–42. URL: http://www.jstor.org/stable/24953849.

Hawking, S. W. and G. Ellis (1973). *The Large Scale Structure of Space-Time.* Cambridge Monographs on Mathematical Physics. Cambridge University Press.

Lovelace, Geoffrey et al. (Oct. 2008). "Binary-black-hole initial data with nearly extremal spins". In: *Phys. Rev. D* 78 (8), p. 084017. DOI: 10.1103/PhysRevD.78.084017. URL: https://link.aps.org/doi/10.1103/PhysRevD.78.084017.

Manchak, John Byron (2013). "Global space time structure". In: *The Oxford Handbook of Philosophy of Physics.* Ed. by Robert Batterman. DOI: 10.1093/oxfordhb/9780195392043.013.0017.

Matzner, Richard A. (1968). "Almost Symmetric Spaces and Gravitational Radiation". In: *Journal of Mathematical Physics* 9.10, pp. 1657–1668. DOI: 10.1063/1.1664495. eprint: https://doi.org/10.1063/1.1664495. URL: https://doi.org/10.1063/1.1664495.

Maudlin, Tim, Elias Okon, and Daniel Sudarsky (2020). "On the status of conservation laws in physics: Implications for semiclassical gravity". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 69, pp. 67–81. ISSN: 1355-2198. DOI: https://doi.org/10.1016/j.shpsb.2019.10.004. URL: https://www.sciencedirect.com/science/article/pii/S1355219819300772.

McMullin, Ernan (1985). “Galilean idealization”. In: *Studies in History and Philosophy of Science Part A* 16.3, pp. 247–273. ISSN: 0039-3681. DOI: https://doi.org/10.1016/0039-3681(85)90003-2. URL: https://www.sciencedirect.com/science/article/pii/0039368185900032.

Menon, Tarun and Craig Callender (2013). “Turn and Face the Strange... Ch-Ch-Changes: Philosophical Questions Raised by Phase Transitions”. In: *The Oxford Handbook of Philosophy of Physics*. Ed. by Robert W. Batterman. Oxford University Press.

Misner, Charles W., K. S. Thorne, and J. A. Wheeler (1973). *Gravitation*. San Francisco: W. H. Freeman. ISBN: 978-0-7167-0344-0, 978-0-691-17779-3.

Noether, Emmy (1918). “Invariante Variationsprobleme.” In: *Nachr. v. d. Ges. d. Wiss. zu Göttingen*. English translation by M.A. Tavel, Reprinted from ”Transport Theory and Statistical Mechanics” 1(3), 183-207 (1971), pp. 235–257.

Norton, John D. (2012). “Approximation and Idealization: Why the Difference Matters”. In: *Philosophy of Science* 79.2, pp. 207–232. DOI: 10.1086/664746. eprint: https://doi.org/10.1086/664746. URL: https://doi.org/10.1086/664746.

— (2016). “The impossible process: Thermodynamic reversibility”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 55, pp. 43–61. ISSN: 1355-2198. DOI: https://doi.org/10.1016/j.shpsb.2016.08.001. URL: https://www.sciencedirect.com/science/article/pii/S1355219815300563.

Page, Don N (Sept. 2005). “Hawking radiation and black hole thermodynamics”. In: *New Journal of Physics* 7, pp. 203–203. DOI: 10.1088/1367-2630/7/1/203. URL: https://doi.org/10.1088/1367-2630/7/1/203.

Palacios, Patricia (2019). “Phase Transitions: A Challenge for Intertheoretic Reduction?” In: *Philosophy of Science* 86.4, pp. 612–640. DOI: 10.1086/704974.

Potochnik, Angela (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Read, James (2020). “Functional Gravitational Energy”. In: *The British Journal for the Philosophy of Science* 71.1, pp. 205–232. DOI: 10.1093/bjps/axx048.

Rosen, Joseph (2008). "Symmetry Rules: How Science and Nature are founded on Symmetry". In.

Schwarzschild, Karl (1916). "On the gravitational field of a mass point according to Einstein's theory". In: *Sitzungsber.Preuss.Akad.Wiss.Berlin (Math.Phys.)* Translated by S. Antoni and A. Loinger (1999). Available at https://arxiv.org/abs/physics/9905030., pp. 189–196.

Turck-Chièze, S (Jan. 2016). "The Standard Solar Model and beyond". In: *Journal of Physics: Conference Series* 665.1, p. 012078. DOI: 10.1088/1742-6596/665/1/012078. URL: https://dx.doi.org/10.1088/1742-6596/665/1/012078.

Turck-Chièze, Sylvaine and Sébastien Couvidat (Aug. 2011). "Solar neutrinos, helioseismology and the solar internal dynamics". In: *Reports on Progress in Physics* 74.8, p. 086901. DOI: 10.1088/0034-4885/74/8/086901. URL: https://dx.doi.org/10.1088/0034-4885/74/8/086901.

Wald, Robert M. (2001). "The Thermodynamics of Black Holes". In: *Living Reviews in Relativity* 4.6.

Zurek, W. H. and Kip S. Thorne (May 1985). "Statistical Mechanical Origin of the Entropy of a Rotating, Charged Black Hole". In: *Phys. Rev. Lett.* 54 (20), pp. 2171–2175. DOI: 10.1103/PhysRevLett.54.2171. URL: https://link.aps.org/doi/10.1103/PhysRevLett.54.2171.

Chapter 5

The Time in Thermal Time

5.1 Preamble: the problem of time

In the Hamiltonian approach to quantum gravity, we formulate general relativity in Hamiltonian form with appropriate constraints and quantize – just as we obtained quantum mechanics from classical mechanics. However, dramatically, the resulting Wheeler-DeWitt equation appears absent of dynamical evolution. Schematically:

$$\hat{H}|\Psi\rangle = 0 \tag{5.1}$$

where $\hat{H}$ is the Hamiltonian operator, and $|\Psi\rangle$ is the quantum state associated with the wave-function(-al) Ψ representing both matter content and geometry in the universe. Importantly, the geometries here are *not* 4-dimensional objects, but 3-dimensional spatial slices.

While the usual Schrödinger equation

$$\hat{H}|\Psi\rangle = i\hbar\frac{\partial|\Psi\rangle}{\partial t} \tag{5.2}$$

describes time-evolution of $|\Psi\rangle$, the Wheeler-DeWitt equation – its quantum gravity analogue – does not. Time 'disappears'. Assuming this equation describes the fundamental state of affairs, the ontology of quantum gravity appears fundamentally timeless: the fundamental ontology – what there ultimately is – contains no reference to time or concepts depending on time. Yet, it seems manifest that our familiar physical systems are – or at least appears to be – evolving *in time*. We are therefore confronted with the problem of recovering time-evolution from fundamentally timeless ontology: the *problem of time*.[1]

Decades of research have ensued, with many attempts to solve, resolve, or dissolve the problem of time. Here, I assess one such proposal: the *thermal time hypothesis*, due originally to Connes & Rovelli (1994). In their own words:

> A radical solution to this problem [...of this absence of a fundamental physical time at the ... generally covariant level...] is based on the idea that one can extend the notion of time flow to generally covariant theories, but this flow depends on the thermal state of the system ... the notion of time flow extends naturally to generally covariant theory, provided that:
>
> 1. we interpret the time flow as a one-parameter group of automorphisms of the observable algebra (generalized Heisenberg picture);
> 2. we ascribe the temporal properties of the flow to thermodynamical causes, and therefore we *tie the definition of time to thermodynamics* and
> 3. we take seriously the idea that in a generally covariant context *the notion of time is not state-independent, as in non-relativistic physics, but rather depends on the state in which the system is in.* (Connes & Rovelli 1994, 2901, emphasis mine)

I take these to be core tenets of the thermal time hypothesis: despite the problem of time and timeless context, time can emerge due to thermodynamic origins. In particular, if we can find systems in special thermodynamic states within the fundamentally timeless ontology – *Kubo-Martin-Schwinger (KMS) thermal states* – then the proposal comes in three parts: first, we use these states to define a privileged one-parameter group of automorphisms given a certain algebraic structure, second, we interpret this group dynamically as a bona fide time parameter, and third, we explain this in terms of thermodynamic considerations. This seems to me to be the natural reading of this hypothesis.[2]

[1] This is actually just *one* of many problems of time. See Kuchar (1991), Isham (1993), Kuchar (2011), Anderson (2017), or Thébault (2021).

[2] Rovelli (personal correspondence) has since distanced himself from this stronger hypothesis, suggesting that it merely provides analysis of 'timely' notions associated with thermodynamics, but *doesn't define* time using thermodynamics. Importantly, he claims that the hypothesis already assumes that some notion of time (perhaps

Prima facie, the thermal time hypothesis is an elegant solution to the problem of time. It's commonly accepted that systems are in thermal equilibrium if certain thermodynamic parameters are unchanging over time, but the hypothesis reverses this observation by proposing instead that the time parameter *defined by* systems in states of thermal equilibrium (thermal states) – specifically, states satisfying the KMS condition. Furthermore, *every* state of interest uniquely picks out some such parameter. This offers an attractive solution to the problem of time requiring only thermodynamic reasoning: if we can find systems in thermal states within the fundamentally timeless ontology, we recover (thermal) time. This makes the hypothesis worthy of investigation in itself.

To my knowledge, philosophers of physics have not dedicated much attention to the hypothesis, beyond Swanson (2014 ch. 5, and 2021).[3] Thus, I hope to fill this lacuna. While Swanson (2021) focuses largely on technical issues with the thermal time hypothesis, I focus on two major conceptual issues. As such, I hope my discussion can complement Swanson's, and generate more interest in the thermal time hypothesis as a solution to the problem of time.

In what follows, I argue that the thermal time hypothesis falls short of solving the problem of time: it's either circular or remains yet unjustified in its derivation of time from no-time through thermodynamic considerations. I'll press my concerns 'from above' and 'from below': from thermodynamic considerations, and from fundamental considerations about the algebraic structure itself.

From 'above', the parameter – specifically, the parameter featuring in the one-parameter group of automorphisms – with respect to which statistical states satisfy the KMS condition cannot always be justifiably interpreted as time because it need *not* align with the system's actual dynamics. Justifying the alignment requires already grasping and identifying systems in thermal equilibrium. However, there appears to be no adequate definition for thermal

defined via relational clocks) is defined. This weaker thermal time hypothesis *does* avoid the conceptual problems I'll raise. However, it's contrary to a natural reading of the above passage. I'll set this weaker thermal time hypothesis aside – it already assumes some (presently absent) resolution of the problem of time by assuming that time can be defined.

[3] Earman (2011, fn. 6) notes that "a discussion of this fascinating proposal will be reserved for another occasion."

equilibrium that doesn't depend on time.

From 'below', an appropriate physical interpretation of the algebraic structure – necessary for understanding thermal states – requires time in at least two ways: First, the way one typically interprets *expectation values* in statistical mechanics, in connecting fundamental physics to thermodynamics, requires an understanding of *fluctuation*. Second, the property of *unitarity* is baked into the algebraic structure used to justify the interpretation of some parameter as thermal time. However, I argue that both notions of fluctuation and unitarity require a background notion of time to acquire physical meaning.

In closing, I discuss a generalization of the thermal time proposal, the *modular time hypothesis*: what if we jettison thermodynamics and appeal directly to the algebraic structure of quantum physics? I argue that similar problems 'from below' persist, even if problems 'from above' disappear. The thermal time hypothesis must address these formidable problems before it can be a plausible solution to the problem of time.

5.2 The time from thermal time

To understand the thermal time hypothesis, I'll first introduce the technical machinery in terms of which the hypothesis is cast.

5.2.1 The Heisenberg picture

The thermal time hypothesis is cast in terms of the Heisenberg picture of quantum mechanics. In standard physics, the Heisenberg picture takes states $|\Psi\rangle$ in Hilbert space $\mathcal{H}$ to be time-*in*dependent. What evolves unitarily over time are the time-*dependent* observables $\mathcal{O}$ – representing measurement outcomes of physical quantities one might make on systems – which are represented by self-adjoint linear operators $A \in \mathcal{A}$ acting on $\mathcal{H}$. However, the physical interpretation of $\mathcal{A}$ is similar to those in the Schrödinger picture: they are the possible (measurement outcomes for) physical quantities attributable to systems at a time.[4]

$\mathcal{A}$ evolves unitarily in accordance with the Heisenberg equation of motion:

$$\frac{\partial A}{\partial t} = \frac{i}{\hbar}[H, A] \tag{5.3}$$

and:

$$A(t) = U^{\dagger}(t)\, A(0)\, U(t) \tag{5.4}$$

where $U(t) = e^{-iHt/\hbar}$ is the unitary operator, $[\cdot, \cdot]$ is the commutator defined as $[A, B] = AB - BA$, and H is the Hamiltonian operator.

5.2.2 C^* algebra: from abstract to concrete

The Heisenberg picture lends naturally to an *algebraic* interpretation of quantum mechanics. The structure of how $\mathcal{A}$ can act on $\mathcal{H}$ can be understood in terms of abstract algebra: it's a non-commutative C^* algebra, specifically, a von Neumann algebra.[5] Furthermore, this abstract algebraic structure of C^* algebras can be represented in terms of the familiar Hilbert space structure as corresponding *concrete* C^* algebras.

Given an abstract C^* algebra $\mathcal{C}$,[6] one can define *states* ω over $\mathcal{C}$: bounded, normalized, positive linear functionals such that

$$\omega : \mathcal{C} \rightarrow \mathbb{C} \tag{5.5}$$

Given that it's normalized and positive, we may also understand $\omega(\mathcal{C})$ as assigning *expectation values* to the elements of $\mathcal{C}$, the physical quantities $\mathcal{A}$. (I return to this point in §5.3.2.)

[4] In generally covariant contexts, Rovelli & Smerlak (2011, 6) suggests: "in quantum gravity the pure states can be given by the solutions of the Wheeler-DeWitt equation, and observables by self-adjoint operators on a Hilbert space defined by these solutions." Of course, such a Hilbert space remains elusive.

[5] For technical exposition, see Bratteli & Robinson (1987 and 1997). For exposition targeted at philosophical audiences, see Ruetsche (2011a, Ch. 4).

[6] An *abstract C^** algebra is a set $\mathcal{C}$ of elements – such as the observables we are interested in – which satisfies various formal properties: it's closed under addition, scalar multiplication, (non-commutative) operator multiplication, involution operation *, and equipped with a norm $|| \; ||$ satisfying $||A^*A|| = ||A||^2$ and $||AB|| \leq ||A||\,||B||$ for all $A, B \in \mathcal{C}$.

Now we connect these abstract notions to concrete physics in terms of Hilbert space structure: this means finding concrete counterparts to the algebraic structure and states via what is called a *representation.* Notably, for each ω on $\mathcal{C}$, the Gelfand-Naimark-Segal (GNS) construction provides a representation $\pi_\omega(\mathcal{C})$ of $\mathcal{C}$ in some Hilbert space $\mathcal{H}_\omega$.[7] Within $\mathcal{H}_\omega$, we are provided with a cyclic and separating vector $|\Psi_\omega\rangle \in \mathcal{H}_\omega$ such that:[8]

$$\omega(A) = \langle\Psi_\omega|\pi(A)|\Psi_\omega\rangle \tag{5.6}$$

In Ruetsche's words, "the expectation value the state ω assigns the algebraic element A is duplicated by the expectation value the vector $|\Psi_\omega\rangle$ assigns to the Hilbert space operator $\pi(A)$". (2011a, 92) This shows that every abstract ω has concrete counterparts as operators on $\mathcal{H}$. Furthermore, one can always find states which guarantee a *faithful* representation; these are *faithful* states.[9] Roughly, faithful representations preserve the abstract algebraic structure in the concrete setting.[10]

These results establish correspondence between the abstract states and algebraic structure of $\mathcal{C}$ and a concrete representation in terms of bounded linear operators A acting on $\mathcal{H}$. We'll call (the norm-closed sub-algebra of) the algebra of bounded linear operators $\mathcal{A}$ acting on $\mathcal{H}$ *concrete* C^* algebras. We furthermore focus on concrete *von Neumann* algebras $\mathcal{W}$ per Connes & Rovelli.[11]

[7] A *representation* is a *-homomorphism $\pi : \mathcal{D} \to \mathcal{B}(\mathcal{H})$ where $\mathcal{D}$ is some abstract algebra, and $\mathcal{B}(\mathcal{H})$ is the set of bounded linear operators on $\mathcal{H}$.

[8] A vector $|\Psi\rangle \in \mathcal{H}$ is *cyclic* for $\mathcal{C}$ just in case $\mathcal{W}|\Psi\rangle$ is dense in $\mathcal{H}$. $|\Psi\rangle \in \mathcal{H}$ is said to be *separating* for $\mathcal{C}$ just in case $A|\Psi\rangle = 0$ implies $A = 0$ for any $A \in \mathcal{C}$.

[9] See Feintzeig (2023, §2.3). Swanson (2021, 286) notes that faithful states must be *mixed* states; nontrivial C^* algebras have no pure, faithful states. Swanson worries that this might force upon the thermal time hypothesis an ignorance interpretation. However, as he notes, Wallace (2012) argues that mixed states needn't require an ignorance interpretation. See also Chen (2021).

[10] Technically, faithful states ensure that π is a $^*-$homomorphism to a subset of bounded operators on $\mathcal{H}$. Such states satisfy the condition that $\omega(A^*A) = 0$ entails $A = 0$ for all $A \in \mathcal{C}$.

[11] These are concrete C^* algebras closed under the weak operator topology satisfying $\mathcal{W} = \mathcal{W}''$. For an algebra $\mathcal{D}$ of bounded operators on $\mathcal{H}$, its commutant $\mathcal{D}'$ is the set of all bounded operators on $\mathcal{H}$ which commutes with every element of $\mathcal{D}$. If $\mathcal{D}$ is an algebra, so is $\mathcal{D}'$. $\mathcal{D}''$ is the *double commutant*, the set of all bounded operators in $\mathcal{D}'$ commuting with $\mathcal{D}'$.

Relative to this concrete structure, Connes & Rovelli restricts their attention to states ω over $\mathcal{W}$ which are represented in $\mathcal{H}$ as *normal states*: normed bounded positive linear functionals satisfying countable additivity.[12] Via the GNS construction, normal states are represented as trace-class density operators ρ with $Tr(\rho) = 1$ in $\mathcal{H}$:

$$\omega(A) = Tr(\rho A) \tag{5.8}$$

for all $A \in \mathcal{W}$, viz., states ω admit density operator representations in $\mathcal{H}$. While Connes & Rovelli don't elaborate much on this restriction, this restriction is presumably motivated by a demand that the algebraic formalism be given clear physical meaning.[13] After all, (9) recovers the standard way of deriving expectation values – observed statistics – of measurements associated with observables $\mathcal{A}$:

$$\langle A \rangle = Tr(\rho A) \tag{5.9}$$

Presumably, any representation should recover this relation. We'll see a worry with this demand for a physical interpretation of expectation values in §5.3.2.

5.2.3 From kinematics to possible dynamics

So far we've focused on algebraically representing the kinematics of standard quantum theory – the structure of $\mathcal{A}$ acting on $\mathcal{H}$ and their expectation values – *at a time*. But the algebraic structure of $\mathcal{W}$ also provides us with something that *could* be understood as time-evolution. Crucially, *any faithful, normal* ω defines a *unique* 1-parameter group of automorphisms of $\mathcal{W}$:

$$\alpha_t^\omega : \mathcal{W} \to \mathcal{W} \tag{5.10}$$

[12] For any countable set of pairwise orthogonal projection operators $\{E_i\} \in \mathcal{W}$:

$$\sum_i \omega(E_i) = \omega(\sum_i E_i) \tag{5.7}$$

[13] Ruetsche (2011b) details why we should demand normal states.

for real t. Given a concrete $\mathcal{W}$ defined by some faithful, normal ω via the GNS construction, the Tomita-Takesaki theory provides a unique α_t in terms of two modular invariants generated from the adjoint conjugation operation *. The theory guarantees the existence of a well-defined operator S:

$$SA|\Psi\rangle = A^*|\Psi\rangle \tag{5.11}$$

and that S uniquely provides a polar decomposition:[14]

$$S = J\Delta^{1/2} \tag{5.12}$$

where J is antiunitary and Δ is a self-adjoint positive operator. α_t, associated with the defining state ω, is then defined by:

$$\alpha_t^\omega A = \Delta^{it} A \Delta^{-it} \tag{5.13}$$

and this uniquely defines a strongly continuous 1-parameter unitary group of automorphisms on $\mathcal{W}$, parametrized by $t \in \mathbb{R}$, which is also called the *modular group*. Crucially, α_t^ω acquires dynamical meaning because Δ^{-it} can be interpreted as unitary operators. With this interpretation, (5.13) should look familiar: it's equivalent to the Heisenberg equation (5.4). The crucial step, in interpreting the modular group defined by any faithful, normal state dynamically, is to interpret its parameter t as playing *the same role as physical time in unitary evolution* in the usual Heisenberg equation. (More on this in §5.3.2.)

However, even if we *can* interpret α_t^ω dynamically, these are very *special* dynamics: any faithful, normal state ω (or the associated ρ) is invariant under the 'flow' of α_t^ω:

$$\omega(\alpha_t^\omega A) = \omega(A) \tag{5.14}$$

Interpreted dynamically, α_t^ω leaves ω unchanged over time. But, importantly, systems in an

[14] See Takesaki (1970).

arbitrary state need *not* be in an unchanging state over time; in these cases, the dynamics associated with α_t^ω doesn't describe the dynamics of that system. Put simply: the two notions of dynamics – the 'dynamics' of α_t^ω and the system's actual dynamics – don't always 'line up' and we are *not* always justified to interpret α_t^ω as the actual dynamics of that system. (More on this in §5.3.1.)

5.2.4 Justifying a dynamical interpretation: from thermal states to KMS states

Importantly, there is one clear case when we *are* physically justified in interpreting α_t^ω dynamically. This is when a system is in *thermal equilibrium*, i.e. a thermal state. In such cases, we are looking for a description of a state that is time-translation-invariant (i.e. stationary), satisfying certain thermodynamic properties e.g. being at constant temperature. In standard physics, this notion is defined in terms of some background time and dynamics. Then, *given* the special kind of dynamics associated with such a state – dynamics which doesn't change the system's thermodynamic state over time – the associated modular group automorphisms α_t^ω can *then* be directly interpreted as the actual dynamics for systems in such a state. For this special case, the dynamics associated with α_t^ω seems to 'line up' with the dynamics of a system in thermal equilibrium.

When do we know that a state ω is thermal? It turns out that one can understand thermal states (with inverse temperature β, $0 < \beta < \infty$) as states satisfying the KMS condition, i.e. as KMS states.[15] KMS states satisfy the following conditions: for any $A, B \in \mathcal{W}$, there exists a complex function $F_{A,B}(z)$, analytic in the strip $\{z \in \mathbb{C} \mid 0 < \text{Im } z < \beta\}$ and continuous on the boundary of the strip, such that for all $t \in \mathbb{R}$:

$$F_{A,B}(t) = \omega(\alpha_t^\omega(A)B) \tag{5.15}$$

[15] $\beta = \frac{1}{k_b T}$, where T is the system's temperature, and k_b is Boltzmann's constant.

$$F_{A,B}(t+i\beta) = \omega(B\alpha_t^\omega(A)) \tag{5.16}$$

$$\omega(\alpha_t^\omega(A)B) = \omega(B\alpha_t^\omega(A)) \tag{5.17}$$

The KMS condition is seemingly arcane, but a crucial physical anchor for this condition – and *why* we can interpret states satisfying this condition as states in thermal equilibrium – is the fact that imposing this condition on a state is formally equivalent, in the finite-dimensional case, to demanding the state is the Gibbs state ρ_β. This is the quantum generalization of the statistical state for systems describable with the grand canonical ensemble, one in thermal equilibrium at constant inverse temperature β with Hamiltonian H:

$$\rho_\beta = \frac{e^{-\beta H}}{Tr(e^{-\beta H})} \tag{5.18}$$

For any operator $A \in \mathcal{A}$, the expectation value for that observable for this system is:

$$\langle A \rangle_\rho = Tr(\rho_\beta A) = \frac{Tr(e^{-\beta H}A)}{Tr(e^{-\beta H})} \tag{5.19}$$

ρ_β satisfies the KMS condition. Interpreting ω in terms of ρ_β via (5.19), and $\alpha_t^\omega(A)$ as $e^{iHt}Ae^{-iHt}$ per (5.4), we get, for operators $A, B \in \mathcal{W}$:

$$Z^{-1}Tr(e^{-\beta H}e^{iHt}Ae^{-iHt}B) = Z^{-1}Tr(e^{-\beta H}e^{-iH(t+i\beta)}Be^{iH(t+i\beta)}A) \tag{5.20}$$

where $Z = Tr(e^{-\beta H})$ is the partition function of ρ_β.[16] From (5.20) we can see that ρ_β satisfies the KMS condition (5.17). Note, again, that we must first justify interpreting α_t^ω dynamically as the unitary operator of the Heisenberg equation!

For finite-dimensional quantum systems, ρ_β uniquely describes systems satisfying the KMS condition.[17] This imbues the seemingly purely syntactic KMS condition (as Emch & Liu

[16] This uses the fact that the Hamiltonian commutes with itself, and that the trace is cyclic.

[17] See Emch & Liu (2002, 351–352). In infinite-dimensional quantum systems, the trace is ill-defined, and so likewise for ρ_β defined using the trace. Crucially, the KMS condition can still hold.

(2002, 351) describes it) with physical meaning, motivating the *physical equivalence* of states formally satisfying the KMS condition and thermal states. Furthermore, KMS states satisfy stability and passivity conditions we typically associate with thermal states.[18]

Importantly, *any* ω satisfies the KMS condition relative to the modular group defined by itself, i.e. α_t^ω,[19] for $\beta = 1$.[20] On a naïve reading, this seems to overgeneralize: *any* state *is* a KMS state, even the state of my cup of coffee which is clearly cooling down. However, the appropriate reading is that any state *can* be a KMS state: there are *some possible* dynamics for a system which keeps it in thermal equilibrium. This is just to reiterate that α_t^ω can but needn't necessarily be interpreted dynamically. It needn't align with a system's actual dynamics. (I'll elaborate in $\S 5.3.1$.)

Returning to the question of when to interpret α_t^ω dynamically, it seems that we are justified to do so when we are justified to interpret a system as being in thermal equilibrium. *If* we know that the system's dynamics – associated with being in thermal equilibrium – 'lines up' with the sort of special dynamics described by α_t^ω, *then* we can interpret α_t^ω dynamically.

5.2.5 The thermal time hypothesis

So far I've introduced everything in a standard quantum mechanical context, where there is assumed to be some background time. In the timeless context, there is no time with which we may determine a system to be in thermal equilibrium, and hence no straightforward way to understand the physical meaning of KMS states (and hence to understand its associated modular group dynamically).

The thermal time hypothesis reverses this situation. Instead of defining thermal equi-

[18] We've already seen that ω is invariant under the flow of α_t^ω; interpreted as a dynamical flow, it captures the idea that an equilibrium state is stationary and doesn't change over time. Some other examples: such a state should not change in free energy over time (thermodynamic stability), remains (over time) in a stationary state arbitrarily close to it under small perturbations (dynamical stability), and energy cannot be extracted from such a state by applying for any finite amount of time any local perturbation of the dynamics α (passivity). See Emch & Liu (2002, 355).

[19] See Bratteli & Robinson (1997).

[20] Ruetsche (2011a, Ch. 7, fn. 23) notes that a state satisfying the KMS condition for $\beta = 1$ also satisfies it for arbitrary $\beta > 0$.

librium, KMS states, and the modular group *in terms of time*, Connes & Rovelli hypothesizes that we simply *define* the modular group to *be* time. Note that this implicitly assumes the applicability of the C^* algebraic structure even in the timeless context.

The motivation for the hypothesis stems from the aforementioned fact that any faithful, normal, state ω defines a preferred one-parameter group of automorphisms α_t^ω. *If* we are further justified in interpreting ω as describing the same physical situation described by the Gibbs state ρ_β (or its KMS generalization), i.e. as a thermal state, *then* we can interpret the dynamics of α_t^ω as being generated by a 'thermal' Hamiltonian $H = -ln\rho_\beta$.[21] Note that the Hamiltonian is defined *in terms of the Gibbs state.* This is contrary to the usual understanding where the Hamiltonian, interpreted prior to the Gibbs state, is used to define the Gibbs state. In other words, given a thermal state, a Hamiltonian can be extracted from it in non-generally covariant contexts. The further claim of the thermal time hypothesis is that we can do *the same thing* in generally covariant contexts *as well.*

To sum up the hypothesis: in the generally covariant context of quantum gravity, where the problem of time looms, we appeal to the C^* algebraic structure and hypothesize that *the flow of time is defined by the unique 1-parameter state-dependent modular automorphism group* α_t^ω, and dynamical equations can be defined in terms of this flow (e.g. in terms of the Hamiltonian as seen above). Systems in thermal equilibrium thus define time in this fundamentally timeless setting, providing a path forward to tackling the problem of time.

5.3 The time in thermal time

While the technical details of the previous subsection are daunting, the conceptual point is a simple one: in the context of standard quantum mechanics and classical thermodynamics, we always have a background time parameter t, with which we can define dynamical notions such as the notion of equilibrium, time-evolution, stationarity, and so on. However, in the generally covariant setting we don't have access to such a time parameter. But, *if* we had

[21] For more details, see Paetz (2010, §4.2 and §5.2).

access to the structure of $\mathcal{W}$, then any faithful, normal, state ω over $\mathcal{W}$ defines a modular group according to which the KMS condition is satisfied. Connes & Rovelli's proposal is that we first interpret these states ω as equilibrium states, then interpret their dynamics α_t^ω as equilibrium dynamics independent of time. Time is defined *in terms of* the equilibrium dynamics via α_t^ω.

Swanson (2021) points out two technical challenges for this program.[22] Firstly, it's unclear whether the thermal time defined here is capable of recovering proper time in general relativity in more physically realistic settings: observers only observe thermal time matching up to proper time in special cases such as when the observer is uniformly accelerating in flat spacetime described by the vacuum state of a quantum field theory. It remains unclear whether this correspondence generalizes. (I return to this in $\S5.4$ when discussing the 'modular time hypothesis'.) Secondly, it's unclear whether the notion of thermal time has a classical counterpart in the classical limit, though as Swanson (2021, $\S4$) argues, choosing the Poisson algebra – rather than commutative von Neumann algebras – as the appropriate classical counterpart to noncommutative von Neumann algebras allays that problem.

Here, beyond these technical challenges, I'll emphasize two *conceptual* challenges for the proposal. Essentially, the thermal time hypothesis tries to define time *in terms of* the modular group of statistical states satisfying the KMS condition which can be interpreted as states in thermal equilibrium, by working with the C^* algebraic structure. To avoid circularity, and to be a genuine solution to the problem of time, the thermal time hypothesis itself had better not depend on time. If thermal equilibrium and the algebraic structure require interpretation *in terms of* time in order to be justifiably applied and physically meaningful, they would only serve to define time insofar as time has already been defined – a circularity *par excellence*. Furthermore, it would not be a solution to the problem of time, for the problem is precisely that we have no time to begin with.

[22] See also Paetz (2010, Ch. 7) for a more expansive discussion of these challenges.

5.3.1 From above: the time in equilibrium

The first conceptual challenge, then, for the thermal time hypothesis is the provision of a time-independent account of thermal equilibrium.

The time in standard accounts of equilibrium

For Connes & Rovelli, "an equilibrium state is a state whose modular automorphism group is the time translation group" for the non-generally covariant context (1994, 2909), and the hypothesis asserts that this carries over to the generally covariant context as well. However, when are we allowed to interpret the modular automorphism group as the time translation group? As I've already emphasized in the previous section, the modular group flow α_t^ω is a very special sort of dynamics for the associated ω, and cannot be interpreted dynamically automatically. Earman & Ruetsche echo this concern: "the modular group determined by an arbitrary faithful normal state on a von Neumann algebra may lack a natural dynamical interpretation, in which case scare quotes should be understood when referring to β as the inverse temperature." (2005, 570) That is, we are not entitled to interpret *any* (faithful normal) state satisfying the KMS condition as being in equilibrium, or having (equilibrium) thermodynamic properties, without further justification. As emphasized, even in the usual contexts, α_t^ω might not line up with the actual dynamics of a system, and hence might not be interpretable dynamically. We need some further *physical argument* for why a system in some arbitrary state ought to be interpreted as having the dynamics associated with α_t^ω, for why this system's dynamics 'lines up' with that associated with α_t^ω.

This point – that α_t^ω doesn't automatically mandate a dynamical interpretation and prior determination of equilibrium is required – was already emphasized in Haag et al's (1967) landmark paper which first connected the KMS condition to the thermodynamic notion of equilibrium:

> We assumed the existence of an automorphism $A \to A_t$ for which $\omega(A)$ is invariant. It then follows that there exists a unitary operator $U(t) = e^{-iHt/\hbar}$ on

> $\mathcal{H}$, which implements this automorphism. *This does not mean, however, that the system actually moves according to this automorphism.* It only means that it's possible to choose the dynamics, i.e. the interparticle forces and the external forces, such that with these forces the system in the state $\omega(A)$ would be in equilibrium. If the forces happen to be different, the automorphism $A \rightarrow A_t$ is not a time translation, H is not the Hamiltonian of the system and the state $\omega(A)$ is not stationary. (1967, 235, emphasis mine)

Put another way, we are justified in taking the automorphisms seriously as dynamics only when we already have some *prior* determination that the system is *already* in thermal equilibrium, for instance, if we already know the Hamiltonian of the system. Once we have done that, we are justified in describing the system as being in a state satisfying the KMS condition, with its dynamics described by the modular automorphisms. Likewise, Swanson (2021, 12) points out:

> Any statistical state determines thermal dynamics according to which it is a KMS state, however, if ρ is a non-equilibrium state, the resultant thermal time flow does not align with our ordinary conception of time. By the lights of thermal time, a cube of ice in a cup of hot coffee is an invariant equilibrium state! The same problem arises in the quantum domain – only for states which are true equilibrium states will the thermal time correspond to physical time.

In other words, for the thermal time hypothesis to take off, it must rule out the fact that any arbitrary state can define *some* 'thermal time'. It must restrict the hypothesis only to the physically meaningful thermal times defined by a privileged class of states over $\mathcal{W}$, which are 'really' equilibrium states. After all, as I've emphasized, genuinely thermal states are the only states for which α_t^ω aligns with the actual dynamics of the system.

But which states are 'really' equilibrium states? The usual way (in standard quantum mechanical contexts) of picking out equilibrium states refer to thermodynamic properties such as the stationarity, stability, and passivity of systems (in particular, of their macroscopic properties). Emch & Liu (2002, 355) observes that the various thermodynamic stability and passivity conditions associated with some state satisfying the KMS condition typically requires that the state "is assumed tacitly to be stationary with respect to a specified dynamics α".[23] In other words, the judgment of whether a state is in a bona fide equilibrium state appears to *all*

[23] See fn. 20 for some of these conditions.

be defined implicitly in terms of some background time parameter.

More generally, the meaning of 'equilibrium' appears intrinsically dependent on the notion of time. As Callen puts it emphatically in his well-known textbook: "in all systems there is a tendency to evolve toward states in which the properties are determined by intrinsic factors and not by previously applied external influences." These are the equilibrium states, which are "by definition, *time independent*" (1985, 13) such that "the properties of the system must be independent of the past history" (1985, 14).[24] In other words, it appears almost *a priori* that the notion of equilibrium is dependent on the notion of a background time, along which processes evolve, properties cease to change, and states terminate in staticity and quiescence.

However, in the present context with its problem of time, we cannot simply claim that these properties of equilibrium obtain. Furthermore, if equilibrium is defined in terms of these properties, and thermal time requires the concept of equilibrium, then thermal time hasn't really solved the problem of time, since it requires time to take off! *Any* (faithful, normal) state can be deemed to be an 'equilibrium' state with respect to the modular group automorphism's 'thermal time', but this renders the meaning of equilibrium arbitrary. Instead, we would need to provide some story for why a state is 'really' in thermal equilibrium. Such a story is typically provided with respect to some background time, and so it remains unclear what a story might look like, which doesn't refer to time at all in defining notions fixing the concept of equilibrium, such as stationarity or passivity.

We thus run into our first dilemma for the thermal time hypothesis: either thermal time hypothesis is circular, since it implicitly requires a background time parameter. Or it's unjustified, since it has yet to justify why some states are privileged equilibrium states, with which we can cash out the thermal time hypothesis. Without such a story the hypothesis is

[24] Other textbooks make similar claims about the temporal and dynamical nature of equilibrium. Buchdahl (1966) defines equilibrium in terms of staticity – a lack of change over relevant timescales. Landau & Lifshitz (1980) point out how equilibrium states are states which are necessarily arrived at *after some relaxation time*. Caratheodory's (1909) discussion of equilibrium also focuses on relaxation time. Similarly, Schroeder (2021, 2) introduces thermal equilibrium as such: "After two objects have been in contact *long enough*, we say that they are in thermal equilibrium." Matolcsi (2004) conceptualizes equilibrium in terms of the *standstill* property, where a process is standstill when they are not varying in time and have vanishing dynamical quantities.

rendered arbitrary.

The time in timeless equilibrium

In response to this problem, Paetz (2010, §7.6) suggests that we would need some *intrinsic* definition of equilibrium – one that doesn't refer to time – if the thermal time hypothesis is to take off. To my knowledge, timeless definitions of equilibrium are not readily available; the only one of note is due to Rovelli (1993).

We can see how Rovelli's 'timeless' definition of equilibrium is supposed to work through classical statistical mechanics. Rovelli (1993, 1559) claims that this condition was emphasized in Landau & Lifshitz's (1980) textbook as a *definition* of equilibrium.[25] For a system S with coordinates p, q, such that we can separate a small but macroscopic region S' with associated phase space coordinates p', q', from the (much larger) rest of the system S'' with phase space coordinates p'', q'', assuming weak interactions between S' and S'', the interaction Hamiltonian approximately vanishes. (See 5.1.) *As a result*, for such a choice of S' and S'', the probability distribution for the system – its statistical state ρ – factorizes:

$$\rho(p,q) = \rho'(p',q')\rho''(p'',q'') \tag{5.21}$$

This condition essentially signals the statistical independence of one sub-system's statistical state from the other. One way to interpret this statistical independence is to understand it as representing a system being in equilibrium with itself by representing its parts (i.e. subsystems) as being in *relative equilibrium with each other.* If these subsystems are in equilibrium with each other, their thermodynamic properties will, of course, not change with respect to each other, and so it might seem natural that the subsystem statistical states – which determine macroscopic measurable (e.g. thermodynamic) quantities – are independent of each other and will factorize. Rovelli then proposes that this condition alone can define equilibrium states: "we

[25] To my knowledge, Landau & Lifshitz does *not* use this condition as a definition of equilibrium, but as a property which (more or less) holds for equilibrium systems.

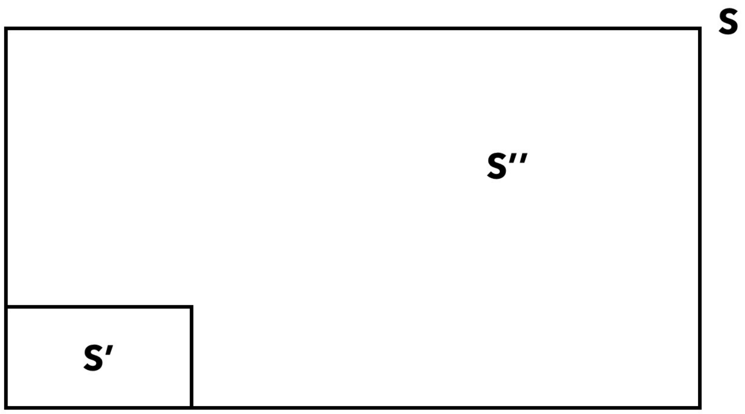

Figure 5.1. A partition of a system S into two subsystems S' and S''.

shall refer to equilibrium as a situation in which every small but still macroscopic component of the system is in equilibrium, in the usual sense, with the rest of the system." (1993, 1558–1559)

We can break down Rovelli's proposed timeless definition into two parts. Firstly, a system is in equilibrium if and only if *every* subsystem is in relative equilibrium with the rest of the system, viz. S' and S'' are in relative equilibrium for all choices of S' and S'' such that S' and S'' are still macroscopic regions and S' is significantly smaller than S''. Secondly, two subsystems S' and S'' are in relative equilibrium if and only if (5.21) holds.[26]

Unfortunately, I don't think that this defintion of equilibrium is adequate. To begin with, the original physical justification for applying (5.21) appears to rely implicitly on time, even if its form is explicitly timeless. We can see this by looking back to Landau & Lifshitz's (1980) introduction of the factorization condition. What Rovelli does *not* mention is Landau & Lifshitz's caveat which immediately precedes (5.21):

> It should be emphasised once more that this property holds only over *not too long intervals of time*. Over a sufficiently long interval of time, the effect of interaction of subsystems, however weak, will ultimately appear. Moreover, it is just this relatively weak interaction which leads finally to the establishment of statistical equilibrium. (1980, 6, emphasis mine)

In other words, the application of this definition is manifestly justified in terms of a background

[26] Landau & Lifshitz (1980, 7) notes that groups of subsystems also factorize with respect to the rest of the system in the same way, provided that these groups are still small enough relative to the rest of the system.

dynamics, just like other definitions of equilibrium. (5.21) is clearly not intended to *define* equilibrium states. Rather, the subsystems of systems in equilibrium can be justifiably characterized in terms of (5.21) *for suitable periods of time* but not *always*. Relative to some timescales, subsystem interactions may be taken to approximately vanish. Over long enough periods of time, interactions between subsystems, however small, will render (5.21) false. Macroscopic properties not changing for subsystems of a system in equilibrium does *not* mean that their probability distributions, which depend on *microphysical properties*, are likewise independent of each other. For all practical purposes, we may treat (5.21) as approximately true, since we typically don't deal with systems on those time-scales. As a conceptual point, though, (5.21) only holds true relative to certain timescales, and should not be taken to be a definition of equilibrium.

One possible response is that we can simply take Landau & Lifshitz's definition but reject their physical justification for this definition. After all, they are clearly not working in the timeless generally covariant context, so, prima facie, we should not expect their justification to apply in this new context.[27] Instead, we should treat (5.21), the factorization of statistical states, to define equilibrium for a generally covariant quantum system. Insofar as systems (approximately) factorize this way, we can take them to be in equilibrium and to define thermal time. We should therefore treat Landau & Lifshitz's original physical justification – that systems factorize *because* they are weakly interacting subsystems in relative thermal equilibrium – as a *consequence* of this definition instead. *Because* of this definition – because systems approximately factorize in this way – we can *then* treat its subsystems as weakly interacting in relative equilibrium with each other.

Nevertheless, I think that both parts of Rovelli's proposed definition face conceptual worries. Firstly, defining equilibrium in terms of relative equilibrium for *all* choices of S' and S'' is too strong. While it's true that a system would be in equilibrium if each subsystem is in such a relative equilibrium with the rest of the system, I don't think that the latter is *necessary*,

[27] I thank an anonymous reviewer for suggesting this worry.

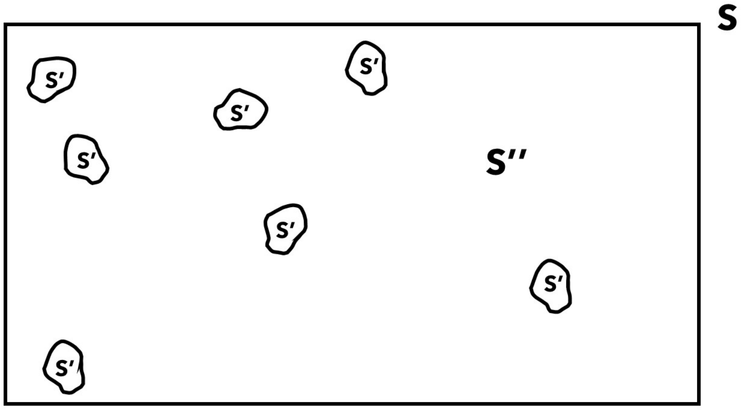

Figure 5.2. A schematic demonic partition of a system S into two subsystems S' and S'', where S' picks out a composite subsystem of higher mean kinetic energy, and S'' picks out a region of lower mean kinetic energy.

and hence cannot be definitional, for equilibrium. Even in the standard non-generally covariant context, any typical system, even those which we *do* know to be in equilibrium, will fail to satisfy the criterion of relative equilibrium for *all* choices of subsystems. The only requirement proposed by Rovelli is that S' and S'' are macroscopic subsystems, and that S' is much smaller than S''. Landau & Lifshitz (1980, 7) notes that the same relation holds for groups of subsystems so long as the group remains small relative to the rest of the system. However, without further constraints, there is always a gerrymandered 'Maxwell's demon' partition of the system into two subsystems:[28] a small (possibly disconnected) collection of subsystems containing all and only the faster particles with higher momentum, S_{fast}, and a much larger region of the system containing all and only the slower particles with lower momentum S_{slow}.[29] (See 5.2.) It seems to me that nothing rules out the possibility of partitioning the system this way. It then follows that S_{fast} is at a much higher temperature than S_{slow} since the former has higher mean kinetic energy. So it seems that relative equilibrium doesn't obtain for such a partition of subsystems, even though we know that the system is in equilibrium overall.

[28] Note that this need *not* be an actual partition (using walls, membranes, etc.), and so sidesteps the question of whether Maxwell's demon is physically realizable.

[29] See e.g. Hemmo & Shenker (2010). Many thanks to Craig Callender and Eddy Keming Chen for suggesting this example in personal correspondence.

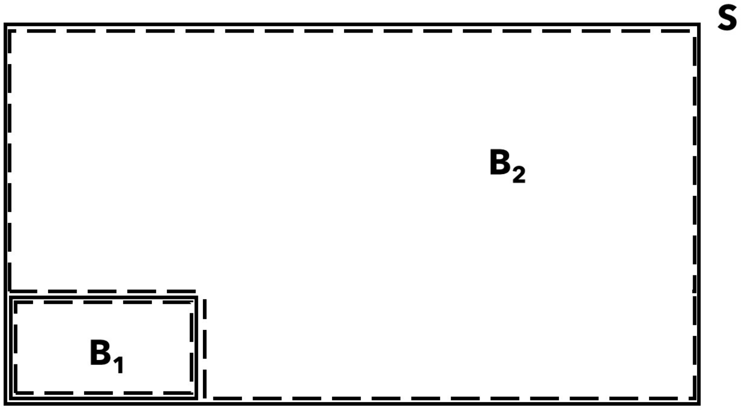

Figure 5.3. A system S containing two (approximately) non-interacting boxes B_1 and B_2, such that the temperature of B_1 is not equal to the temperature of B_2. Their states factorize since they are non-interacting, but they are not in relative thermal equilibrium.

Secondly, there's a worry about whether (5.21) – the factorization condition – suffices to track whether two subsystems are in relative equilibrium. Consider the simple case of a system of two boxes, B_1 and B_2. (See 5.3.) B_2 can be much larger than B_1. The boxes are thermally insulated, electromagnetically shielded, and contains air at different temperatures. It seems to me that we can ascribe a statistical state ρ_B to the joint system of B_1 and B_2, and that, at least for some regimes,[30] ρ_B is factorizable into two subsystem statistical states ρ_{B_1} and ρ_{B_2}, each describing the statistical state of the respective boxes. Taken as is, the proposed 'timeless' definition of relative equilibrium appears to hold. However, it does *not* suffice to characterize these two boxes as actually being in relative equilibrium: the two boxes are, ex hypothesi, at different temperatures.

Note that the above problems would not be troubling if we took the proposed condition to hold *over time* and allowed the subsystems to *interact*, viz. if the definition were justified by appeal to a background time. Then S_{fast} would quickly lose energy to S_{slow} and equilibrate over time. But this would require some notion of time, as with Landau & Lifshitz's original

[30] ρ_B might be factorizable into ρ_{B_1} and ρ_{B_2} simpliciter if we have perfect thermal insulation and perfect mirrors preventing the transmission of radiation. Otherwise, there'll be some regimes for which we can ignore thermal radiation especially if the two boxes are far apart, and hence regimes for which we might plausibly assume factorizability. The point is that, for whatever regime in which the factorizability condition holds, there is no clear physical sense in which the two systems are in relative equilibrium because they are not at the same temperature.

physical justification. For Landau & Lifshitz, (5.21) is justified when we are allowed to take a system to be comprised of *quasi-closed* systems – when these subsystems *interact weakly* with each other. (1980, Ch. 1, §2) In the original context, the statistical independence encapsulated in (5.21) seems to be justified in terms of (approximate) *lack of interaction.* But interaction appears to be something steeped in *dynamics* and thus time. It doesn't seem to be something we have without time. Since we are not allowed this justificatory resource, we are not allowed this natural solution to the problem raised in the previous paragraph.

Furthermore, the problem is exacerbated in the timeless context. Note that Rovelli's definition was proposed against the background of classical mechanics, where the relevant states of the system in terms of p, q are defined *at a time*. In the timeless context, the problem of time precludes a similar restriction of states to those at the same time, since time is exactly what's missing. As a result, the insufficiency of factorization to define equilibrium is readily amplified. Even without going into the timeless state space of quantum gravity (something that remains out of reach), we can see how the insufficiency of factorization is magnified when we don't restrict attention to states at a time. Consider the application of the proposed factorization condition to a system S undergoing a probabilistic process such that at each time step τ, the state of the system at τ is either 1 or 0 with some probability. Furthermore, the state is probabilistically independent of future and past outcomes. Then, for any arbitrary partition of the entire sequence *across* time into sub-sequences, the two 'subsystems' factorize and so satisfy the timeless definition of relative equilibrium. (See 5.4.) But it's clear that these subsystems are *not* in relative thermal equilibrium – they are simply probabilistically independent of each other.

To sum up, Rovelli's 'intrinsic' definition of equilibrium was originally justified by Landau & Lifshitz with respect to some background time and dynamics. In the timeless context, we can of course reject Landau & Lifshitz's justification on account of irrelevance. But now we have no clear justification for the definition. Furthermore, the definition appears to be inadequate for defining equilibrium states: relative equilibrium for all choices of partitions is

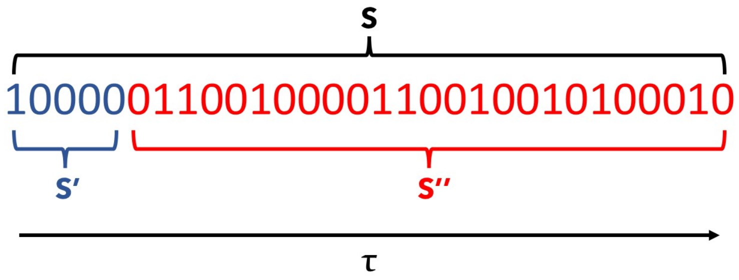

Figure 5.4. A probabilistic process with outcomes $\{0, 1\}$ each with some probability of occurring every time step τ. Outcomes at each time step are probabilistically independent of future and past outcomes. The probability distribution for the sequence S clearly factorizes for any choice of sub-sequences S', S''.

unnecessary for characterizing equilibrium, and factorization is insufficient for characterizing relative equilibrium.

Without an unproblematic 'timeless' definition of equilibrium, though, we remain unable to pick out privileged ω over $\mathcal{W}$ that we can use to define thermal time, without circularity. Thus, the dilemma – surrounding how to appropriately define equilibrium in the context of quantum gravity – remains for the thermal time hypothesis.

5.3.2 From below: the time in algebraic structure

However, *even if* we settled the prior conceptual challenge related to understanding equilibrium, I argue that the thermal time hypothesis still has a formidable challenge 'from below': justifying the physical interpretation of the algebraic structure without appealing to time.[31]

This second conceptual challenge is related to whether two core concepts – expectation values and unitarity – employed by the algebraic structure can be interpreted in a physically meaningful way without a background time to begin with. Without time, it seems that we

[31] See Unruh (1997) for similar, though more general, considerations.

cannot justify a physical interpretation of the formal expectation values assigned by ω, nor can we interpret α_t^ω dynamically without already interpreting the unitary group in terms of time. And without these concepts, we cannot connect ω to thermal states. Hence the challenge ‘from below’.

The time in physical expectation values

Recall that statistical states of interest – density operators corresponding to faithful normal states ω over $\mathcal{W}$ – are defined as mappings of elements in $\mathcal{W}$ to complex numbers. From §5.2 we saw that they *can* be interpreted as representing the *expectation value* of an observable A via:

$$\omega(A) = Tr(\rho A) \tag{5.22}$$

due to its properties of being real, positive and normalized, thus lending itself to a probabilistic interpretation. This then, I claimed, connects these statistical states to the physical quantities of actual physical processes. But how *does* it do so in ordinary, non-generally covariant cases?

In classical statistical mechanics, statistical states are given physical meaning in terms of their expectation values which are given by:

$$\langle f \rangle = \int f(p, q)\rho_{cg}(p, q)dpdq \tag{5.23}$$

where f is any physical quantity dependent on canonical momenta and positions p and q, and ρ_C is the classical probability density function.

Interpreting $\langle f \rangle$ rests on what Frigg & Werndl (2021) call the ‘averaging principle’: “what we observe in an experiment on a system is the ensemble average of a phase function representing a relevant physical quantity.” But how *do* we physically interpret a probabilistic, averaged quantity? Furthermore, in the standard Gibbs framework, this average is a property of *ensembles* of systems, rather than of any single system in particular. However, as Landau & Lifshitz (1980, 4) point out, an expectation value – a “statistical averaging” over ensembles of systems in their words – “is exactly equivalent to a time averaging”, specifically, averaged over time as time approaches infinity. One imagines a trajectory that traverses all of available phase phase such that “during a sufficiently long time” the system will “be in many times in every possible state”. (Landau & Lifshitz 1980, 3) Over long enough times, the probability of a system being found in some state (i.e. some region of phase space) is equivalent to the volume of said region. As Frigg & Werndl (2021) point out, this is the ‘standard’ textbook justification of the averaging principle. An alternative interpretation is to interpret the expectation values as a value around which measurements *at a time* might vary, i.e. *fluctuate*, such that we might expect such fluctuations to vanish *over time*.[32] For both cases, we seem to already presuppose some sort of dynamics, in order for us to interpret the expectation values physically and ascribe properties to the system in question.

In ordinary quantum mechanics, the situation is not too different. As Isham (1993, 246) puts it, one typically interprets a density matrix ρ as “an ensemble of systems in which every element possesses a value for [observable] A, and the fraction having the value a_i is $|\psi_i|^2$; i.e., an essentially classical probabilistic interpretation of the results of measuring A”, such that:

$$\rho = |\psi_i|^2 |a_i\rangle\langle a_i| = |\psi_i|^2 A_i \tag{5.24}$$

where A_i is the projection operator associated with state a_i and $|\psi_i|^2$ is the probability given by the wave-function ψ that a system has property a_i at a time, in accordance with the Born rule. The probability for observing a system in state a_i, as seen before, is $Tr(\rho A_i)$. Again, though, this expectation value is a property of an ensemble of systems. How are we supposed to interpret it for single systems?

As with the classical case before, Landau & Lifshitz (1980) propose an interpretation of the expectation value (“mean value” in their terms) for single systems in terms of *fluctuations* over time: for any physical quantity of interest for a single system, “*in the course of time*

[32] See Frigg & Werndl (2021).

this quantity varies, fluctuating about its mean value." (1980, 7, emphasis mine) For large macroscopic thermodynamic systems, fluctuation terms approximately vanish, and hence "the quantity itself may be regarded as practically constant *in time* and equal to its mean value." (1980, 9) They emphasize in particular that the fact that "the relative·fluctuations of additive physical quantities tend to zero as the number of particles increases made no use of any specific properties of classical mechanics, and so remains entirely valid in the quantum case." (1980, 19) Again, the connection between the statistical state and physical systems, through the expectation values, appears to require some theory of fluctuations, i.e. change *over time*.

Returning to the thermal time hypothesis, recall that it postulates the applicability of the structure of von Neumann algebras $\mathcal{W}$ even in the timeless context. Then, states ω defined on $\mathcal{W}$ are interpreted as thermal states in virtue of satisfying the KMS condition with respect to α_t^ω. α_t^ω is then interpreted as thermal time. But to interpret ω as picking out genuine physical states in our world means to interpret their associated expectation values as physical quantities. But without a background time to begin with, how are we to interpret these expectation values as fluctuations? With respect to *what* are they fluctuating? Without such a story, it seems to me that we cannot have a physical interpretation of ω in the timeless context, and so the program doesn't appear to take off to begin with.

One possible response might be to assume that they are not fluctuating with respect to time, but with respect to 'modal space' (e.g. the as-of-yet-unknown Hilbert space of the Wheeler-DeWitt equation). But that just takes us back to the old question in classical statistical mechanics, of how phase averages – fluctuations of quantities in 'modal space' i.e. ensembles of systems – could tell us anything about the *actual* physical quantities of the actual world.

The time in unitarity

The second and perhaps more concerning problem with applying the C^* algebraic structure in the timeless context concerns the fact that the thermal time hypothesis's use of this structure appears to depend on some physically meaningful notion of unitarity. Formally, it's an operator that conserves the inner product of Hilbert space, quantities of the form $\langle\psi_1|A|\psi_2\rangle$ for states ψ_1, ψ_2, and operators A. Unitarity, in my view, is the glue that connects the abstract algebraic structure to a dynamical interpretation, which is required for interpreting α_t^ω as a genuine dynamical object. This, in turn, is required to interpret α_t^ω as thermal time. However, in the timeless context, this property is not guaranteed to obtain at all.

Recall my comment from the end of §5.2.3: the Tomita-Takesaki theory guarantees the existence of a unique strongly continuous 1-parameter unitary group of automorphisms on $\mathcal{W}$, the modular group, parametrized by $t \in \mathbb{R}$ such that:

$$\alpha_t^\omega A = \Delta^{-it} A \Delta^{it} \tag{5.25}$$

But what is the *physical meaning* of the modular group? The crucial point was to interpret Δ^{it} as a unitary operator. It was this step that conceptually connected (5.25) to physics via the Heisenberg equations of motion, which allowed us to interpret α_t^ω as a bona fide dynamical object. Furthermore, the connection between states satisfying the KMS condition and the more familiar Gibbs states, discussed in §5.2.4, also required us to interpret α_t^ω as unitary evolution. In the ordinary quantum context and even in quantum field theory, this is of course not a problem, since we have some background (space-)time with which one may define unitary operators.

However, in the context of the problem of time and a fundamentally timeless ontology, it's not at all given that there *is* a way to implement unitarity, and hence a way to interpret α_t^ω dynamically in terms of unitarity. If there's no such interpretation available, we no longer appear to have a justification for interpreting α_t^ω as a dynamical object. Indeed, we've so far discussed the thermal time hypothesis *as though* there already is a physically meaningful background Hilbert space with a conserved inner product. As a matter of mathematics, of course, once we start with an abstract algebra of observables, we automatically get these structures for free via the GNS representation and so on. However, the question is whether the mathematics has any physical meaning. While we can readily interpret the formal unitary operators physically in ordinary quantum mechanical cases, with respect to some background time, it's not clear at all that a fundamentally timeless context immediately supports such an interpretation. In that case, it's not clear that these formal structures have any physical meaning at all. It then becomes unclear whether the algebraic structure is the right kind of structure to apply in the timeless context.

In the specific case of the Wheeler-DeWitt equation, Kuchar (2011, 9) notes that "nothing similar is granted by the Wheeler-DeWitt equation". More generally, beyond that context *per se*, this is the well-known and yet-unresolved *Hilbert space problem*, the problem of defining an conserved positive-definite inner product on Hilbert space *without time*.[33] The task is to endow on the space of solutions to the Wheeler-DeWitt equation a conserved inner product along some parameter and complete it into a Hilbert space. This task is what Kuchar (2011) calls the Hilbert space problem. However, as Kuchar (40-42) explains, the inner product one naturally defines on this space of solutions is *not* positive-definite. One can avoid this problem only if the Hamiltonian of the Wheeler-DeWitt equation satisfies certain conditions. But these conditions are generically *not met*, and formidable technical challenges await: the Hamiltonian is not stationary, the potential term in the Hamiltonian needn't be positive, and we cannot rule out states with negative energy as unphysical. So we can't find a parameter along which unitarity holds. Furthermore, even if we *could* find a conserved inner product in this space of solutions, the connection between this inner product and the usual inner product and its conservation in spacetime is still opaque (Kuchar (2011, 42) calls this the 'spacetime problem'). In short, technical challenges prevent the implementation of unitarity in the timeless context of the Wheeler-DeWitt equation. Without unitarity, though, it's not clear *why* we are allowed to apply the C^* algebraic structure *and* interpret it physically, e.g. interpret α_t^ω as a dynamical object.

So the second challenge for the thermal time hypothesis is this: find a notion of unitarity

[33] See Kuchar (2011) or Anderson (2017).

that doesn't depend on time, with which we may then interpret α_t^ω, from which time can finally be defined.

One might object that I am demanding too much.[34] Given our limited state of understanding of quantum gravity, perhaps it seems fair for theorists to postulate what sorts of structures might be present at the fundamental level in quantum gravity, and then try to see whether time shows up *given the existence of such structures.* I agree that this is generically an acceptable strategy, but I have two reservations in the context of the thermal time hypothesis.

Firstly, we can of course postulate the existence of whatever structures we want, but I maintain that we need reason to expect such structures to obtain. In the timeless context of the Wheeler-DeWitt equation, there is yet no clear notion of unitarity, nor a time along which we can understand expectation values. Since these are core concepts employed by the C^* algebraic structure (in order to give the abstract mathematics a physical interpretation), it seems fair for me to question how the proponent of thermal time might justify the existence of such algebraic structure.

Secondly, it seems to me that the obtaining of this structure, in requiring unitarity (or at least something like it), renders the thermal time hypothesis inert: the *thermal* part might be rendered irrelevant. After all, if we *could* find some notion of unitarity even at the most basic level of the algebraic structure – define some inner product which is conserved over some parameter – and we accept that this parameter is time in standard quantum mechanics, then it seems to me that there would already be some notion of time present, *prior* to thermodynamic considerations: it's the notion of time along which probability currents are conserved. So if we *could* interpret α_t^ω dynamically in terms of unitarity, we already have time and the thermal part of the thermal time hypothesis is irrelevant. If we couldn't, then time doesn't exist 'all the way up', and the thermal time hypothesis doesn't take off. Either way, it spells trouble for the thermal time hypothesis.

[34] I thank an anonymous reviewer for suggesting this worry.

5.4 The modular time hypothesis: away with thermodynamics?

The last discussion suggests an interesting alternative. In response to my second reservation, one might reply: so much the worse for the thermal time hypothesis! If we could solve the problem of time by demanding the existence of such an algebraic structure at the fundamental timeless level, we have no need for the thermal time hypothesis anyway.

Instead, we could postulate something like the 'modular time hypothesis'.[35] On this hypothesis, we forsake the dependence on thermodynamics and define time *directly in terms of the modular group*. Stated more precisely, for any system in a faithful, normal state ω, the associated α_t^ω *is* time.

There are at least two reasons to consider this hypothesis in lieu of the thermal time hypothesis. Firstly, the worry 'from above' raised in §5.3.1 suggests that interpreting the modular group dynamics as equilibrium dynamics requires an antecedent notion of thermal equilibrium, which is conceptually challenging. Jettisoning the requirement that time emerges 'from thermodynamical considerations', might be the more felicitous thing to do.

Secondly, there are well-known results due to Bisognano & Wichmann (1975, 1976) which suggest a direct connection between the modular group and spacetime geometry outside the quantum gravity context, *without* the need for thermodynamic considerations. Rather, the thermodynamic interpretation becomes a *downstream effect* of this connection between the modular group and spacetime geometry. More specifically, given a Minkowski vacuum state over the Weyl algebra $\mathcal{A}(\mathbb{R}^4)$ of the Klein-Gordon field and the associated von Neumann algebra $\mathcal{W}(\mathcal{O})$ associated with an open region of spacetime $\mathcal{O}$, the restriction of the algebra to the right Rindler wedge $\mathcal{R}$ (see 5.5.) leads to a geometrical interpretation for the associated modular group for the Minkowski vacuum state: its generators are the Lorentz boosts on $\mathcal{R}$.[36]

[35] I borrow this from Earman's (2011) discussion of the 'modular temperature hypothesis'. I thank an anonymous reviewer for suggesting this line of thought.

[36] See Earman (2011) and Swanson (2021).

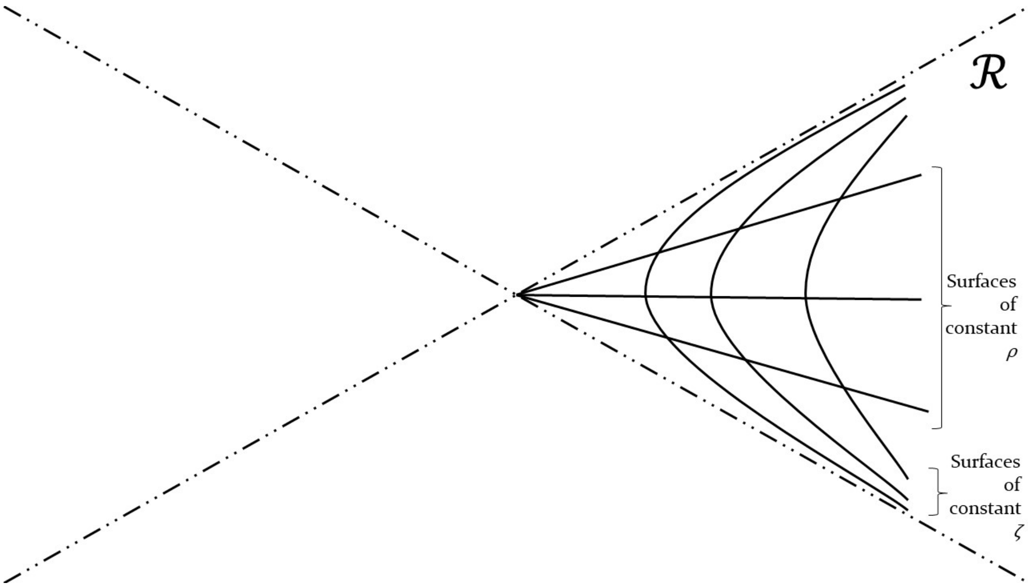

Figure 5.5. Schematic representation of the right Rindler wedge $\mathcal{R}$ in Minkowski spacetime. The Rindler coordinates $\{\zeta, y, z, \rho\}$ are related to the Minkowski coordinates $\{x, y, z, t\}$ by $x = \zeta\text{cosh}\rho$ and $t = \zeta\text{sinh}\rho$ and the Minkowski metric in Rindler coordinates becomes $ds^2 = d\zeta^2 + dy^2 + dz^2 - \zeta^2 d\rho^2$. Surfaces of constant ρ and ζ are labeled, where ρ is 'Rindler time' and the timelike surfaces of constant ζ can be interpreted as the worldlines of constantly accelerating observers. Dashed lines indicate null directions.

As Swanson (2014, 143) explains, given the Bisognano-Wichmann theorem, we can interpret Δ^{it} in terms of boosts $U(t) = e^{2\pi itK_1}$ where K_1 is the representation of the generator of a boost in the ζ-direction.[37] In the right Rindler wedge, it's already known that Lorentz boosts $\Lambda(a\tau)$ implement a proper time translation along the orbit of observers with constant acceleration a, and so $U(\tau) = e^{ai\tau K_1}$. Simply put, in this specific context, we can see how α_t^ω can be interpreted as a dynamical object *without* thermodynamics; it lines up naturally with the proper time of constantly accelerating observers. Furthermore, this connection between modular time and proper time justifies the connection between the modular group and thermal states: the restriction of the vacuum state to $\mathcal{R}$ is a KMS-state relative to the modular group with an Unruh temperature $\frac{a}{2\pi}$.[38] Importantly, here, the alignment of the modular group and proper time is justified not by thermodynamics but by considerations about spacetime geometry. As such,

[37] Symmetries can be represented as unitary operators given Wigner's theorem.

[38] I set c, k and $\hbar$ to 1. See Earman (2011).

this avoids the worry 'from above' I raised. Proponents of modular time might then use this connection in the relativistic context to justify extending this connection even to the timeless context, using the modular group to define time there.

For these two reasons, proponents of the the thermal time hypothesis, faced with the challenge 'from above' with defining thermal equilibrium in a timeless fashion, might be tempted to give up the thermal time hypothesis and simply appeal to the modular time hypothesis instead.

A preliminary note: observe that this proposal is not like the thermal time hypothesis, in that time is *not* emergent. After all, the algebraic structure of the modular group exists 'all the way down', if it exists at all in the timeless context. Once we can find a set of (diffeomorphism-invariant) observables of interest, the algebraic structure – and hence the modular group – can be defined as a matter of mathematics. Contrast this with the thermal time hypothesis, where thermodynamics is almost always understood as an emergent theory, and hence likewise for the associated thermal time.

This is unlikely to be a worry for proponents of modular time. After all, the worry is not about emergence, but about finding time in the timeless context. However, I have two further reservations about this proposal as well.

Firstly, this connection between modular time and proper time exists only for a very specific class of models, and even then for a very specific class of observers in said model: immortally and constantly accelerating observers in Minkowski vacuum. Generally, thermal time and proper time will not line up. Swanson (2021) discusses the technical challenges that arise once we relax these assumptions.[39] If we consider finite observers who has causal access not to the entire right Rindler wedge but a finite causal diamond (the intersection of their future light cone at 'birth' and past light cone at 'death'), the two quantities don't generally converge and hence modular time doesn't have the neat geometrical interpretation we desire. Likewise when we consider nonuniform acceleration and nonvacuum states. Yet the modular

[39] See Swanson (2021, §3).

time hypothesis in the timeless context is intended to recover physical time *in general*, rather than physical time in the special context of Rindler observers. Proponents of modular time must justify why modular time should play the role of physical time outside of these contexts, where there is little evidence to suggest that they play the role of physical time. Furthermore, the same worries that motivated us earlier to search for timeless notions of equilibrium return: generally, the modular group dynamics will not align with the actual dynamics of a given system. Why should we treat it as a dynamical object even in standard cases, let alone the timeless context?

Secondly, the above connection between modular time and proper time is established in the standard relativistic context, where there *is* a background time given to us via the metric. Even if we resolved my first reservation, a further question is whether this connection is expected to hold in the timeless context. Given my earlier worries 'from below', we have reasons to be concerned. Why are we justified to treat α_t^ω as dynamical at all? Even for the modular time hypothesis, we seem to need to first be able to interpret α_t^ω in terms of unitary operators $e^{2\pi itK_1}$. But why are we justified in allowing ourselves the property of unitarity in the timeless context? And if we *could* specify unitarity in the timeless context, why won't the time featuring in *that*, rather than the modular group per se, be time? Furthermore, what do the expectation values of states over the algebraic structure mean, in the timeless context? These all lead us back to the worries 'from below' from the previous section.

As such, even though the modular time hypothesis avoids the worry 'from above' by jettisoning thermodynamics, it still runs into the same worries 'from below'. Furthermore, without the physical meaning provided by thermodynamics, the modular time hypothesis also appears to struggle to justify why modular time lines up with proper time outside of a very special class of models.

5.5 Conclusion

Overall, given the conceptual problems surrounding thermodynamic considerations and the algebraic structure itself in the timeless context, the thermal time hypothesis appears to me to be either circular, or yet unjustified.

Importantly, note that I am *not* committed to the *impossibility* of there being a satisfactory justification period. Rather, I simply want to emphasize that there has *yet* to be a satisfactory justification for applying certain thermodynamic concepts, as well as dynamical concepts in the algebraic structure, in the fundamentally timeless setting. The thermal time hypothesis's appeal to thermodynamic reasoning in the quantum gravity regime is a delicate issue, since the conceptual foundations of thermodynamics beyond classical domains is not entirely secure: even in special relativity, temperature 'falls apart' in the sense that there is no single coherent concept of relativistic temperature to be found.[40] More conceptual groundwork needs to be laid. That is to say, I leave open the possibility that the challenges raised here for the thermal time hypothesis may be met.

In any case, this should make clear just how conceptually challenging the problem of time is for quantum gravity researchers. A similar problem also arises in the semiclassical approach to the problem of time, where time is assumed to emerge from semiclassical approximations. Chua & Callender (2021) argues that these approximations, too, implicitly assume time, and are yet unjustified otherwise. There is "no time for time from no-time", in their words. Likewise, it seems that there's 'no time for thermal time from no-time' for now.

Acknowledgments

"The Time in Thermal Time", in part, is currently under review at *Philosophy of Science*. The dissertation author was the sole author of this paper.

[40] See e.g. Chua (forthcoming).

Bibliography

Anderson, Edward (2017). *Problem of time: Quantum Mechanics Versus General relativity.* Springer.

Bisognano, Joseph J. and Eyvind H. Wichmann (1975). “On the duality condition for a Hermitian scalar field”. In: *Journal of Mathematical Physics* 16.4, pp. 985–1007. doi: 10.1063/1.522605. eprint: https://doi.org/10.1063/1.522605. url: https://doi.org/10.1063/1.522605.

— (1976). “On the duality condition for quantum fields”. In: *Journal of Mathematical Physics* 17.3, pp. 303–321. doi: 10.1063/1.522898. eprint: https://aip.scitation.org/doi/pdf/10.1063/1.522898. url: https://aip.scitation.org/doi/abs/10.1063/1.522898.

Bratteli, Ola and Derek W. Robinson (1987). *Operator Algebras and Quantum Statistical Mechanics 1: C*- and W*-algebras, symmetry groups, decomposition of states.* Springer.

Bratteli and Robinson (1997). *Operator Algebras and Quantum Statistical Mechanics 2: Equilibrium States. Models in Quantum Statistical Mechanics.* 2nd. Springer.

Buchdahl, Hans Adolf (1966). *The Concepts of Classical Thermodynamics.* Cambridge University Press.

Callen, Herbert Bernard (1985). *Thermodynamics and an introduction to thermostatistics.* John Wiley & Sons, Inc.

Chen, Eddy Keming (2021). “Quantum Mechanics in a Time-Asymmetric Universe: On the Nature of the Initial Quantum State”. In: *The British Journal for the Philosophy of Science* 72.4, pp. 1155–1183. doi: 10.1093/bjps/axy068. eprint: https://doi.org/10.1093/bjps/axy068. url: https://doi.org/10.1093/bjps/axy068.

Chua, Eugene Y. S. (2023). “T Falls Apart: On the Status of Classical Temperature in Relativity”. In: *Philosophy of Science* 90.5.

Connes, A and C Rovelli (1994). “Von Neumann algebra automorphisms and time thermodynamics relation in generally covariant quantum theories”. In: *Classical and Quantum Gravity* 11.12, pp. 2899–2917. DOI: 10.1088/0264-9381/11/12/007.

Dirac, P. A. M. (1964). *Lectures on Quantum Mechanics.* Dover Publications.

Earman, John (2011). “The Unruh effect for philosophers”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42.2, pp. 81–97. DOI: 10.1016/j.shpsb.2011.04.001.

Earman, John and Laura Ruetsche (2005). “Relativistic invariance and modal interpretations”. In: *Philosophy of Science* 72.4, pp. 557–583. DOI: 10.1086/505448.

Emch, Gérard G. and Chuang Liu (2002). *The logic of thermostatistical physics.* Springer.

Feintzeig, Benjamin H. (2023). *The Classical–Quantum Correspondence.* Elements in the Philosophy of Physics. Cambridge University Press. DOI: 10.1017/9781009043557.

Frigg, Roman and Charlotte Werndl (2021). “Can somebody please say what Gibbsian Statistical Mechanics says?” In: *The British Journal for the Philosophy of Science* 72.1, pp. 105–129. DOI: 10.1093/bjps/axy057.

Haag, R., N. M. Hugenholtz, and M. Winnink (1967). “On the equilibrium states in Quantum Statistical Mechanics”. In: *Communications in Mathematical Physics* 5.3, pp. 215–236. DOI: 10.1007/bf01646342.

Hemmo, Meir and Orly Shenker (2010). “Maxwell’s demon”. In: *Journal of Philosophy* 107.8, pp. 389–411. DOI: 10.5840/jphil2010107833.

Isham, C. J. (1993). “Canonical Quantum Gravity and the Problem of Time”. In: *Integrable Systems, Quantum Groups, and Quantum Field Theories.* Ed. by L. A. Ibort and M. A. Rodríguez. Dordrecht: Springer Netherlands, pp. 157–287. ISBN: 978-94-011-1980-1. DOI: 10.1007/978-94-011-1980-1_6. URL: https://doi.org/10.1007/978-94-011-1980-1_6.

Kuchar, Karel (1991). “The Problem of Time in Canonical Quantization of Relativistic Systems”. In: *Conceptual Problems of Quantum Gravity.* Ed. by A. Ashtekar and J. Stachel. Birkhauser, pp. 141–171.

— (2011). “Time and interpretations of quantum gravity”. In: *International Journal of Modern Physics D* 20, pp. 3–86. DOI: 10.1142/s0218271811019347.

Landau, L. D. and E. M. Lifshitz (1980). *Statistical physics.* Translated from the Russian by J. B. Sykes and M. J. Kearsley. Pergamon Press.

Matolcsi, Tamas (2004). *Ordinary Thermodynamics.* Akademiai Kiads. ISBN: 9789630581707.

Paetz, Tim-Torben (2010). “An Analysis of the ‘T hermal-Time Concept’ of Connes and Rovelli”. In: Master’s thesis. Georg-August-Universität Göttingen.

Rovelli, C (1993). “Statistical mechanics of gravity and the thermodynamical origin of Time”. In: *Classical and Quantum Gravity* 10.8, pp. 1549–1566. DOI: 10.1088/0264-9381/10/8/015.

Rovelli, Carlo and Matteo Smerlak (2011). “Thermal time and Tolman–Ehrenfest effect: ‘temperature as the speed of time’”. In: *Classical and Quantum Gravity* 28.7, p. 075007. DOI: 10.1088/0264-9381/28/7/075007.

Ruetsche, Laura (2011a). *Interpreting quantum theories.* Oxford University Press.

— (2011b). “Why be normal?” In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42.2. Philosophy of Quantum Field Theory, pp. 107–115. ISSN: 1355-2198. DOI: https://doi.org/10.1016/j.shpsb.2011.02.003. URL: https://www.sciencedirect.com/science/article/pii/S1355219811000153.

Schroeder, D.V. (2021). *An Introduction to Thermal Physics.* Oxford University Press. ISBN: 9780192895547. URL: https://books.google.com/books?id=IRUOEAAAQBAJ.

Swanson, Noel (2014). “Modular Theory and Spacetime Structure in QFT”. Princeton University. PhD thesis.

— (2021). “Can quantum thermodynamics save time?” In: *Philosophy of Science* 88.2, pp. 281–302. DOI: 10.1086/711569.

Takesaki, M. (1970). “Tomita’s theory of Modular Hilbert algebras and its applications”. In: *Lecture Notes in Mathematics*. DOI: 10.1007/bfb0065832.

Thebault, Karim P. Y. (2021). “The Problem of Time”. In: *The Routledge Companion to Philosophy of Physics*. Ed. by Eleanor Knox and Alastair Wilson. Routledge, pp. 386–400.

Unruh, W. G. (1997). “Time, gravity, and quantum mechanics”. In: *Time’s Arrows Today: Recent Physical and Philosophical Work on the Direction of Time*. Ed. by S. F. Savitt. Cambridge University Press, pp. 23–94.

Bibliography

Abramowicz, Marek A. and P. Chris Fragile (Jan. 2013). "Foundations of Black Hole Accretion Disk Theory". In: *Living Reviews in Relativity* 16.1, p. 1. ISSN: 1433-8351. DOI: 10.12942/lrr-2013-1. URL: https://doi.org/10.12942/lrr-2013-1.

Albert, David Z. (2000). *Time and Chance.* Cambridge, Massachusetts: Harvard University Press.

Anderson, Edward (2017). *Problem of time: Quantum Mechanics Versus General relativity.* Springer.

Arnowitt, Richard L., Stanley Deser, and Charles W. Misner (1962). "The Dynamics of general relativity". In: *Gen. Rel. Grav.* 40, pp. 1997–2027. DOI: 10.1007/s10714-008-0661-1. arXiv: gr-qc/0405109.

Balescu, R. (1968). "Relativistic statistical thermodynamics". In: *Physica* 40.3, pp. 309–338. ISSN: 0031-8914. DOI: https://doi.org/10.1016/0031-8914(68)90132-8. URL: https://www.sciencedirect.com/science/article/pii/0031891468901328.

Beetle, Christopher and Shawn Wilder (Mar. 2014). "Perturbative stability of the approximate Killing field eigenvalue problem". In: *Classical and Quantum Gravity* 31.7, p. 075009. DOI: 10.1088/0264-9381/31/7/075009. URL: https://doi.org/10.1088/0264-9381/31/7/075009.

Bekenstein, Jacob D. (1973). "Black Holes and Entropy". In: *Phys. Rev. D* 7.8, pp. 2333–2346. DOI: 10.1103/PhysRevD.7.2333. URL: https://link.aps.org/doi/10.1103/PhysRevD.7.2333.

— (Nov. 1975). "Statistical black-hole thermodynamics". In: *Phys. Rev. D* 12 (10), pp. 3077–3085. DOI: 10.1103/PhysRevD.12.3077. URL: https://link.aps.org/doi/10.1103/PhysRevD.12.3077.

Bisognano, Joseph J. and Eyvind H. Wichmann (1975). "On the duality condition for a Hermitian scalar field". In: *Journal of Mathematical Physics* 16.4, pp. 985–1007. DOI: 10.1063/1.522605. eprint: https://doi.org/10.1063/1.522605. URL: https://doi.org/10.1063/1.522605.

— (1976). "On the duality condition for quantum fields". In: *Journal of Mathematical Physics* 17.3, pp. 303–321. DOI: 10.1063/1.522898. eprint: https://aip.scitation.org/doi/pdf/10.1063/1.522898. URL: https://aip.scitation.org/doi/abs/10.1063/1.522898.

Boltzmann, Ludwig (1896–1995). *Lectures on Gas Theory*. New York: Dover.

Bona, C., J. Carot, and C. Palenzuela-Luque (Dec. 2005). "Almost-stationary motions and gauge conditions in general relativity". In: *Phys. Rev. D* 72 (12), p. 124010. DOI: 10.1103/PhysRevD.72.124010. URL: https://link.aps.org/doi/10.1103/PhysRevD.72.124010.

Bratteli, Ola and Derek W. Robinson (1987). *Operator Algebras and Quantum Statistical Mechanics 1: C*- and W*-algebras, symmetry groups, decomposition of states*. Springer.

Bratteli and Robinson (1997). *Operator Algebras and Quantum Statistical Mechanics 2: Equilibrium States. Models in Quantum Statistical Mechanics*. 2nd. Springer.

Brillouin, Louis Marcel (1956). *Science and Information Theory*. New York: Academic Press.

Brown, Harvey (forthcoming). "Do symmetries "explain" conservation laws? The modern converse Noether theorem vs pragmatism." In: *The Philosophy and Physics of Noether's Theorems*. Ed. by James Read, Bryan Roberts, and Nicholas Teh. Cambridge, UK: Cambridge University Press.

Brush, Stephen (1983). *Statistical Physics and the Atomic Theory of Matter From Boyle and Newton to Landau and Onsager*. Princeton: Princeton University Press.

Bub, Jeffrey (2005). "Quantum mechanics is about quantum information". In: *Foundations of Physics* 35.4, pp. 541–560.

Buchdahl, Hans Adolf (1966). *The Concepts of Classical Thermodynamics*. Cambridge University Press.

Butterfield, J. (2010). "Less is different: Emergence and reduction reconciled". In: *Foundations of Physics* 41.6, pp. 1065–1135. DOI: 10.1007/s10701-010-9516-1.

Butterfield, Jeremy (2011). "Less is Different: Emergence and Reduction Reconciled". In: *Foundations of Physics* 41, pp. 1065–1135.

Callen, Herbert Bernard (1985). *Thermodynamics and an introduction to thermostatistics.* John Wiley & Sons, Inc.

Callendar, Hugh Longbourne (1887). "On the Practical Measurement of Temperature: Experiments Made at the Cavendish Laboratory, Cambridge." In: *Philosophical Transactions of the Royal Society of London* A178, pp. 161–230.

Callender, C. (2001). "Taking thermodynamics too seriously. Studies in History and Philosophy of Science Part B". In: *Studies in History and Philosophy of Modern Physics* 32.4, pp. 539–553. DOI: 10.1016/s1355-2198(01)00025-9.

Callender, Craig (1999). "Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy". In: *The Journal of Philosophy* 96.7, pp. 348–373.

— (2001). "Taking Thermodynamics Too Seriously". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32.4. The Conceptual Foundations of Statistical Physics, pp. 539–553. ISSN: 1355-2198. DOI: https://doi.org/10.1016/S1355-2198(01)00025-9. URL: https://www.sciencedirect.com/science/article/pii/S1355219801000259.

Carathéodory, C. (1909). "Untersuchungen über die Grundlagen der Thermodynamik". In: *Mathematische Annalen* 67.3, pp. 355–386.

Carlip, S. (2014). "Black hole thermodynamics". In: *International Journal of Modern Physics D* 23.11. DOI: 10.1142/S0218271814300237.

Carroll, Sean M. (2019). *Spacetime and Geometry: An Introduction to General Relativity.* Cambridge University Press. DOI: 10.1017/9781108770385.

Chalmers, David (2020). "What is conceptual engineering and what should it be?" In: *Inquiry.* DOI: 10.1080/0020174X.2020.1817141.

Chang, Hasok (2004a). *Inventing Temperature: Measurement and Scientific Progress.* Oxford University Press.

Chang, Hasok (2004b). *Inventing Temperature: Measurement and Scientific Progress*. New York, US: OUP.

Chen, Eddy Keming (2021). "Quantum Mechanics in a Time-Asymmetric Universe: On the Nature of the Initial Quantum State". In: *The British Journal for the Philosophy of Science* 72.4, pp. 1155–1183. DOI: 10.1093/bjps/axy068. eprint: https://doi.org/10.1093/bjps/axy068. URL: https://doi.org/10.1093/bjps/axy068.

Chua, Eugene Y. S. (2021). "Does Von Neumann Entropy Correspond to Thermodynamic Entropy?" In: *Philosophy of Science* 88.1. DOI: 10.1086/710072.

— (2023). "T Falls Apart: On the Status of Classical Temperature in Relativity". In: *Philosophy of Science* 90.5.

Chua, Eugene Y. S. and Craig Callender (2021). "No time for time from no-time". In: *Philosophy of Science* 88.5, pp. 1172–1184. DOI: 10.1086/714870.

Compagner, A. (1989). "Thermodynamics as the continuum limit of statistical mechanics". In: *American Journal of Physics* 57, pp. 106–117. DOI: doi:10.1119/1.16103.

Connes, A and C Rovelli (1994). "Von Neumann algebra automorphisms and time thermodynamics relation in generally covariant quantum theories". In: *Classical and Quantum Gravity* 11.12, pp. 2899–2917. DOI: 10.1088/0264-9381/11/12/007.

Cook, Gregory B. and Bernard F. Whiting (Aug. 2007). "Approximate Killing vectors on S^2". In: *Phys. Rev. D* 76 (4), p. 041501. DOI: 10.1103/PhysRevD.76.041501. URL: https://link.aps.org/doi/10.1103/PhysRevD.76.041501.

Corfield, David (1997). "Assaying Lakatos's philosophy of mathematics". In: *Studies in History and Philosophy of Science* 28, pp. 99–121.

Cubero, David et al. (2007). "Thermal Equilibrium and Statistical Thermometers in Special Relativity". In: *Phys. Rev. Lett.* 99 (17), p. 170601. DOI: 10.1103/PhysRevLett.99.170601. URL: https://link.aps.org/doi/10.1103/PhysRevLett.99.170601.

Curiel, Erik (2019). "The many definitions of a black hole." In: *Nature Astronomy* 3, pp. 27–34.

De Haro, Sebastian (forthcoming). “Noether’s Theorems and Energy in General Relativity”. In: *The Philosophy and Physics of Noether’s Theorems*. Ed. by James Read, Bryan Roberts, and Nicholas Teh. Cambridge, UK: Cambridge University Press.

De Regt, Henk W. (2005). “Kinetic Theory”. In: *The Philosophy of Science: An Encyclopedia*. Ed. by Sahotra Sarkar and Jessica Pfeifer. Routledge.

Denbigh, Kenneth (1990). “How Subjective is Entropy?” In: *Maxwell’s demon: Entropy, Information, Computing*. Ed. by Harvey Leff and Andrew Rex. Princeton, New Jersey.

Deville, Alain and Yannick Deville (2013). “Clarifying the link between von Neumann and thermodynamic entropies”. In: *Eur. Phys. J. H* 38, pp. 57–81. doi: DOI:10.1140/epjh/e2012-30032-0.

Dirac, P. A. M. (1964). *Lectures on Quantum Mechanics*. Dover Publications.

Dougherty, John and Craig Callender (2016). *Black Hole Thermodynamics: More Than an Analogy?* Available at: http://philsci-archive.pitt.edu/13195/. (last accessed 13th March 2022).

Duerr, Patrick M. (Dec. 2019). “Against ‘functional gravitational energy’: a critical note on functionalism, selective realism, and geometric objects and gravitational energy”. In: *Synthese*. issn: 1573-0964. doi: 10.1007/s11229-019-02503-3. url: https://doi.org/10.1007/s11229-019-02503-3.

Dunkel, Jörn, Peter Hänggi, and Stefan Hilbert (Oct. 2009). “Non-local observables and lightcone-averaging in relativistic thermodynamics”. In: *Nature Physics* 5.10, pp. 741–747. issn: 1745-2481. doi: 10.1038/nphys1395. url: https://doi.org/10.1038/nphys1395.

Earman, J. (1974). “An attempt to add a little direction to ”the problem of the direction of Time””. In: *Philosophy of Science* 41.1, pp. 15–47. doi: 10.1086/288568.

Earman, John (1978). “Combining Statistical-Thermodynamics and Relativity Theory: Methodological and Foundations Problems”. In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1978, pp. 157–185.

Earman, John (2011). “The Unruh effect for philosophers”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42.2, pp. 81–97. DOI: 10.1016/j.shpsb.2011.04.001.

Earman, John and John Norton (1998). “Exorcist XIV: The wrath of Maxwell’s Demon. Part I. From Maxwell to Szilard.” In: *Studies in History and Philosophy of Modern Physics* 29, pp. 435–471.

— (1999). “Exorcist XIV: The wrath of Maxwell’s Demon. Part II. From Szilard to Landauer and beyond.” In: *Studies in History and Philosophy of Modern Physics* 30, pp. 1–40.

Earman, John and Miklós Rédei (1996). “Why Ergodic Theory Does Not Explain the Success of Equilibrium Statistical Mechanics”. In: *The British Journal for the Philosophy of Science* 47, pp. 63–78.

Earman, John and Laura Ruetsche (2005). “Relativistic invariance and modal interpretations”. In: *Philosophy of Science* 72.4, pp. 557–583. DOI: 10.1086/505448.

Eddington, A.S. (1928). *The nature of the physical world: Gifford Lectures (1927.* en. The University Press.

Einstein, A. (1979). *Autobiographical Notes.* en. Trans. by P.A. Schilpp. Open Court Printing.

Einstein, Albert (1907). “Uber das Relativitätsprinzip und die aus demselben gezogenen”. In: *Folgerungen. J. Radioakt. Elektron* 4, pp. 411–462.

— (1946/1979). *Autobiographical Notes. (P. A. Schilpp, Trans.)* Open Court Printing.

— (1914/1997). “Contributions to Quantum Theory”. In: *The Collected Papers of Albert Einstein, Volume 6 (English). The Berlin Years: Writings, 1914-1917. (English translation supplement) Translated by Alfred Engel.* Princeton University Press: Princeton, pp. 20–26. DOI: Availableat:https://press.princeton.edu/titles/6161.html.

Emch, Gérard G. and Chuang Liu (2002). *The logic of thermostatistical physics.* Springer.

Farías, Cristian, Victor A. Pinto, and Pablo S. Moya (2017). “What is the temperature of a moving body?” In: *Nature Scientific Reports* 7.1, p. 17657. ISSN: 2045-2322. DOI: 10.1038/s41598-017-17526-4. URL: https://doi.org/10.1038/s41598-017-17526-4.

Feintzeig, Benjamin H. (2023). *The Classical–Quantum Correspondence.* Elements in the Philosophy of Physics. Cambridge University Press. DOI: 10.1017/9781009043557.

Feng, Justin C., Edgar Gasperín, and Jarrod L. Williams (Dec. 2019). "Almost-Killing equation: Stability, hyperbolicity, and black hole Gauss law". In: *Phys. Rev. D* 100 (12), p. 124034. DOI: 10.1103/PhysRevD.100.124034. URL: https://link.aps.org/doi/10.1103/PhysRevD.100.124034.

Fletcher, Samuel C. (June 2014). *Similarity, Topology, and Physical Significance in Relativity Theory.* URL: http://philsci-archive.pitt.edu/17235/.

— (Nov. 2018). "Global spacetime similarity". In: *Journal of Mathematical Physics* 59.11, p. 112501. DOI: 10.1063/1.5052354. URL: https://doi.org/10.1063%2F1.5052354.

— (Mar. 2020b). *Approximate Local Poincaré Spacetime Symmetry in General Relativity.* Forthcoming in Claus Beisbart, Tilman Sauer, and Christian Wüthrich, eds. "Thinking about Space and Time." Einstein Studies, vol. 15, Birkhäuser. URL: http://philsci-archive.pitt.edu/17229/.

— (2020a). "The principle of stability." In: *Philosophers' Imprint* 20.3, pp. 199–220.

Fowler, Michael (n.d.). *Black-Body Radiation.*
URL = https://galileo.phys.virginia.edu/classes/252/black_body_radiation.html (last accessed March 19 2022).

Frigg, Roman and Charlotte Werndl (2021a). "Can Somebody Please Say What Gibbsian Statistical Mechanics Says?" In: *The British Journal for the Philosophy of Science.* DOI: 10.1093/bjps/axy057.

— (2021b). "Can somebody please say what Gibbsian Statistical Mechanics says?" In: *The British Journal for the Philosophy of Science* 72.1, pp. 105–129. DOI: 10.1093/bjps/axy057.

Friston, Karl and Klaas E. Stephan (2007). "Free-energy and the brain". In: *Synthese* 159.3, pp. 417–458.

Gavassino, L. (Dec. 2021). "Proving the Lorentz Invariance of the Entropy and the Covariance of Thermodynamics". In: *Foundations of Physics* 52.1, p. 11. ISSN: 1572-9516. DOI: 10.1007/s10701-021-00518-w. URL: https://doi.org/10.1007/s10701-021-00518-w.

Geroch, Robert (1967). "Topology in General Relativity." In: *Journal of Mathematical Physics* 8, pp. 782–786.

Gibbs, Josiah W. (1902). *Elementary principles of statistical mechanics: Developed with special reference to the rational foundation of thermodynamics.* New Haven: Yale University Press.

Goldstein, Sheldon et al. (2020). "Gibbs and Boltzmann Entropy in Classical and Quantum Mechanics". In: *Statistical Mechanics and Scientific Explanation.* Ed. by Valia Allori. Singapore: World Scientific, pp. 519–581.

Haag, R., N. M. Hugenholtz, and M. Winnink (1967). "On the equilibrium states in Quantum Statistical Mechanics". In: *Communications in Mathematical Physics* 5.3, pp. 215–236. DOI: 10.1007/bf01646342.

Haddad, Wassim M. (2017). "Thermodynamics: The Unique Universal Science". In: *Entropy* 19.11. ISSN: 1099-4300. URL: https://www.mdpi.com/1099-4300/19/11/621.

Hallett, Michael (1979). "Towards a Theory of Mathematical Research Programmes II". In: *The British Journal for the Philosophy of Science* 30.2, pp. 135–159.

— (n.d.). "Towards a Theory of Mathematical Research Programmes I". In: *The British Journal for the Philosophy of Science* 1.1979a (), pp. 1–25.

Haslanger, Sally (2000). "Gender and Race: (What) Are They? (What) Do We Want Them To Be?" In: *Noûs* 34.1, pp. 31–55.

Hawking, S. W. (1974). "Black hole explosions?" In: *Nature* 248, pp. 30–31.

— (1975). "Particle creation by black holes." In: *Commun. Math. Phys.* 43, pp. 199–220.

— (1976). "Breakdown of predictability in gravitational collapse". In: *Phys. Rev. D* 14 (10), pp. 2460–2473.

Hawking, S. W. (1977). “The Quantum Mechanics of Black Holes”. In: *Scientific American* 236.1, pp. 34–42. URL: http://www.jstor.org/stable/24953849.

Hawking, S. W. and G. Ellis (1973). *The Large Scale Structure of Space-Time*. Cambridge Monographs on Mathematical Physics. Cambridge University Press.

Hawking, Stephen (1976). “Black holes and thermodynamics”. In: *Physical Review D* 13.2, pp. 191–197.

Hemmo, Meir and Orly Shenker (2006). “Von Neumann’s Entropy Does Not Correspond to Thermodynamic Entropy”. In: *Philosophy of Science* 73.2, pp. 153–174. URL: http://www.jstor.org/stable/10.1086/510816.

— (2010). “Maxwell’s demon”. In: *Journal of Philosophy* 107.8, pp. 389–411. DOI: 10.5840/jphil2010107833.

Henderson, Leah (2003). “The Von Neumann Entropy: A Reply to Shenker”. In: *British Journal for the Philosophy of Science* 54.2, pp. 291–296. URL: http://www.jstor.org/stable/3541968.

Hughes, R. I. G. (1992). *The Structure and Interpretation of Quantum Mechanics*. Cambridge, Mass: Harvard Univ. Press. DOI: https://doi.org/10.2307/2186092.

Isham, C. J. (1993). “Canonical Quantum Gravity and the Problem of Time”. In: *Integrable Systems, Quantum Groups, and Quantum Field Theories*. Ed. by L. A. Ibort and M. A. Rodríguez. Dordrecht: Springer Netherlands, pp. 157–287. ISBN: 978-94-011-1980-1. DOI: 10.1007/978-94-011-1980-1_6. URL: https://doi.org/10.1007/978-94-011-1980-1_6.

Janis, Allen (2018). “Conventionality of Simultaneity”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2018. Metaphysics Research Lab, Stanford University.

Jaynes, Edward T. (1957). “Information Theory and Statistical Mechanics”. In: *Physical Review* 106.4, pp. 620–630.

Joule, James Prescott and William Thomson (1854/1882). *On the Thermal Effects of Fluids in Motion, Part 2.* In Thomson 1882, 357–400. Originally published in the Philosophical Transactions of the Royal Society of London 144:321–364.

Kampis, George, Ladislav Kvasz, and Michael Stöltzner (2002). *Appraising Lakatos: Mathematics, methodology, and the man.* Dordrecht: Kluwer.

Kiss, Olga (2006). "Heuristic, Methodology or Logic of Discovery? Lakatos on Patterns of Thinking". In: *Perspectives on Science* 14.3, pp. 302–317.

Kuchar, Karel (1991). "The Problem of Time in Canonical Quantization of Relativistic Systems". In: *Conceptual Problems of Quantum Gravity.* Ed. by A. Ashtekar and J. Stachel. Birkhauser, pp. 141–171.

— (2011). "Time and interpretations of quantum gravity". In: *International Journal of Modern Physics D* 20, pp. 3–86. DOI: 10.1142/s0218271811019347.

Lakatos, Imre (1976–2015). *Proofs and refutations: The logic of mathematical discovery.* Ed. by J. Worrall and E. Zahar. Cambridge: Cambridge University Press.

— (1978). *The methodology of scientific research programmes.* Ed. by J. Worrall and G. Currie. Vol. 1. New York: Cambridge University Press.

Landau, L. D. and E. M. Lifshitz (1980). *Statistical physics.* Translated from the Russian by J. B. Sykes and M. J. Kearsley. Pergamon Press.

Landsberg, P. T . (1970). "Special Relativistic Thermodynamics – A Review." In: *Critical Review of Thermodynamics.* Ed. by E. B. Stuart, B. Gal-Or, and A. T. Brainard. Baltimore: Mono Book Corp., pp. 253–72.

Landsberg, P. T. and K. A. Johns (1967). "A relativistic generalization of thermodynamics". In: *Il Nuovo Cimento B (1965-1970)* 52.1, pp. 28–44. DOI: 10.1007/BF02710651.

Landsberg, P. T. and G. E. A. Matsas (2004). "The impossibility of a universal relativistic temperature transformation". In: *Physica A: Statistical Mechanics and its Applications* 340.1, pp. 92–94. ISSN: 0378-4371. URL: https://www.sciencedirect.com/science/article/pii/S0378437104004017.

Landsberg, P.T and K.A Johns (1970). "The Lorentz transformation of heat and work". In: *Annals of Physics* 56.2, pp. 299–318. ISSN: 0003-4916. DOI: https://doi.org/10.1016/0003-4916(70)90020-5. URL: https://www.sciencedirect.com/science/article/pii/0003491670900205.

Landsberg, Peter T. and George E.A. Matsas (1996). "Laying the ghost of the relativistic temperature transformation". In: *Physics Letters A* 223.6, pp. 401–403. ISSN: 0375-9601. DOI: 10.1016/s0375-9601(96)00791-8. URL: http://dx.doi.org/10.1016/S0375-9601(96)00791-8.

Lange, Marc (2002). *An Introduction to the Philosophy of Physics: Locality, Fields, Energy, and Mass.* Blackwell.

Le Chatelier, Henri and O. Boudouard (1901). *High-Temperature Measurements (Trans. George K. Burgess.* New York: Wiley.

Leng, Mary (2002). "Phenomenology and mathematical practice". In: *Philosophia Mathematica* 10, pp. 3–25.

Lewis, M.B. and A.J.F. Siegert (1956). "Extension of the Condensation Theory of Yang and Lee to the Pressure Ensemble". In: *Physical Review* 101.4, pp. 1227–1233.

Liu, Chuang (1992). "Einstein and Relativistic Thermodynamics in 1952: A Historical and Critical Study of a Strange Episode in the History of Modern Physics". In: *British Journal for the History of Science* 25.2, pp. 185–206. DOI: 10.1017/s0007087400028764.

— (1994). "Is There a Relativistic Thermodynamics? A Case Study of the Meaning of Special Relativity". In: *Studies in History and Philosophy of Science Part A* 25.6, pp. 983–1004. DOI: 10.1016/0039-3681(94)90073-6.

Lovelace, Geoffrey et al. (Oct. 2008). "Binary-black-hole initial data with nearly extremal spins". In: *Phys. Rev. D* 78 (8), p. 084017. DOI: 10.1103/PhysRevD.78.084017. URL: https://link.aps.org/doi/10.1103/PhysRevD.78.084017.

Manchak, John Byron (2013). "Global space time structure". In: *The Oxford Handbook of Philosophy of Physics.* Ed. by Robert Batterman. DOI: 10.1093/oxfordhb/9780195392043.013.0017.

Matolcsi, Tamas (2004). *Ordinary Thermodynamics.* Akademiai Kiads. ISBN: 9789630581707.

Matzner, Richard A. (1968). "Almost Symmetric Spaces and Gravitational Radiation". In: *Journal of Mathematical Physics* 9.10, pp. 1657–1668. DOI: 10.1063/1.1664495. eprint: https://doi.org/10.1063/1.1664495. URL: https://doi.org/10.1063/1.1664495.

Maudlin, Tim (2011). *Quantum Non-Locality and Relativity.* John Wiley & Sons, Ltd.

Maudlin, Tim, Elias Okon, and Daniel Sudarsky (2020). “On the status of conservation laws in physics: Implications for semiclassical gravity”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 69, pp. 67–81. ISSN: 1355-2198. DOI: https://doi.org/10.1016/j.shpsb.2019.10.004. URL: https://www.sciencedirect.com/science/article/pii/S1355219819300772.

Maxwell, James Clerk (1878). “Tait’s “Thermodynamics””. In: *Nature* 17, pp. 257–259, 278–280.

McCaskey, John P. (2020). “History of ‘temperature’: maturation of a measurement concept”. In: *Annals of Science* 77.4, pp. 399–444.

McDonald, Kirk (2020). *Temperature and Special Relativity.* (last accessed March 15 2022). URL: http://kirkmcd.princeton.edu/examples/temperature_rel.pdf.

McIrvine, Edward C. and Myron Tribus (Sept. 1971). “Energy and Information”. In: *Scientific American.* URL: https://www.scientificamerican.com/article/energy-and-information/.

McMullin, Ernan (1985). “Galilean idealization”. In: *Studies in History and Philosophy of Science Part A* 16.3, pp. 247–273. ISSN: 0039-3681. DOI: https://doi.org/10.1016/0039-3681(85)90003-2. URL: https://www.sciencedirect.com/science/article/pii/0039368185900032.

Menon, Tarun and Craig Callender (2013). “Turn and Face the Strange... Ch-Ch-Changes: Philosophical Questions Raised by Phase Transitions”. In: *The Oxford Handbook of Philosophy of Physics.* Ed. by Robert W. Batterman. Oxford University Press.

Misner, Charles W., K. S. Thorne, and J. A. Wheeler (1973). *Gravitation.* San Francisco: W. H. Freeman. ISBN: 978-0-7167-0344-0, 978-0-691-17779-3.

Myrvold, Wayne (2011). “Statistical mechanics and thermodynamics: A Maxwellian view”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42.2, pp. 237–243.

Nagel, E. (1968). *The structure of science: Problems in the logic of scientific explanation.* Routledge & Kegan Paul Ltd.

Natsuume, Makoto (2015). *AdS/CFT Duality User Guide in Lecture Notes in Physics 903.* Tokyo: Springer.

Neumann, John von (1955). *Mathematical Foundations of Quantum Mechanics.* Princeton University Press: Princeton.

Nickles, T. (1973). "Two concepts of intertheoretic reduction". In: *The Journal of Philosophy* 70.7, pp. 181–201. DOI: 10.2307/2024906.

Noether, Emmy (1918). "Invariante Variationsprobleme." In: *Nachr. v. d. Ges. d. Wiss. zu Göttingen.* English translation by M.A. Tavel, Reprinted from "Transport Theory and Statistical Mechanics" 1(3), 183-207 (1971), pp. 235–257.

Norton, J.D. (2016a). "The impossible process: Thermodynamic reversibility." In: *Studies in History and Philosophy of Modern Physics* 55, pp. 43–61. DOI: 10.1016/j.shpsb.2016.08.001.

Norton, John (2017). "Thermodynamically reversible processes in statistical physics." In: *American Journal of Physics* 85.135, pp. 135–145.

Norton, John D. (2006). "Atoms, Entropy, Quanta: Einstein's Miraculous Argument of 1905". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 37.1, pp. 71–100. DOI: 10.1016/j.shpsb.2005.07.003.

— (2012). "Approximation and Idealization: Why the Difference Matters". In: *Philosophy of Science* 79.2, pp. 207–232. DOI: 10.1086/664746. eprint: https://doi.org/10.1086/664746. URL: https://doi.org/10.1086/664746.

— (2016b). "The impossible process: Thermodynamic reversibility". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 55, pp. 43–61. ISSN: 1355-2198. DOI: https://doi.org/10.1016/j.shpsb.2016.08.001. URL: https://www.sciencedirect.com/science/article/pii/S1355219815300563.

Ott, H. (1963). "Lorentz-Transformation der Warme und der Temperatur". In: *Zeitschrift fur Physik* 175, pp. 70–104.

Paetz, Tim-Torben (2010). "An Analysis of the 'T hermal-Time Concept' of Connes and Rovelli". In: Master's thesis. Georg-August-Universität Göttingen.

Page, Don N (Sept. 2005). "Hawking radiation and black hole thermodynamics". In: *New Journal of Physics* 7, pp. 203–203. DOI: 10.1088/1367-2630/7/1/203. URL: https://doi.org/10.1088/1367-2630/7/1/203.

Palacios, P. (2018). "Had we but world enough, and time... but we don't!: Justifying the thermodynamic and infinite-time limits in Statistical Mechanics". In: *Foundations of Physics* 48.5, pp. 526–541. DOI: 10.1007/s10701-018-0165-0.

Palacios, Patricia (2019). "Phase Transitions: A Challenge for Intertheoretic Reduction?" In: *Philosophy of Science* 86.4, pp. 612–640. DOI: 10.1086/704974.

Pathria, R K (1966). "Lorentz transformation of thermodynamic quantities". In: *Proceedings of the Physical Society* 88.4, pp. 791–799. DOI: 10.1088/0370-1328/88/4/301. URL: https://doi.org/10.1088/0370-1328/88/4/301.

— (1967). "Lorentz transformation of thermodynamic quantities: II". In: *Proceedings of the Physical Society* 91.1, pp. 1–7. DOI: 10.1088/0370-1328/91/1/302. URL: https://doi.org/10.1088/0370-1328/91/1/302.

Peres, Asher (1990). "Thermodynamic Constraints on Quantum Axioms". In: *Complexity, Entropy, and the Physics of Information, The Proceedings Of The Workshop On Complexity, Entropy and the Physics of Information Held May-June, 1989 in Santa Fe, New Mexico, Santa Fe Institute Studies in the Sciences of Complexity, vol. VIII.* Ed. by Zurek, W. H. Westview Press, pp. 345–356.

— (2002). *Quantum Theory: Concepts and Methods (2nd Ed.)* Dordrecht: Kluwer Academic Publishers.

Planck, M. (1908). "On the dynamics of moving systems". In: *Ann. Phys. Leipz.* 26, pp. 1–34.

— (1920). "The Genesis and Present State of Development of the Quantum Theory". In: *Nobel Lecture, June 2, 1920.* URL: https://www.nobelprize.org/prizes/physics/1918/planck/lecture/.

Planck, M., Verband Deutscher Physikalischer Gesellschaften, and Max-Planck-Gesellschaft zur Förderung der Wissenschaften (1958). *Physikalische Abhandlungen und Vorträge:*

Aus Anlass seines 100. Geburtstages (23. April 1958). Physikalische Abhandlungen und Vorträge. F. Vieweg.

Planck, Max (1908). “Zur Dynamik bewegter Systeme”. In: *Ann. Phys. Leipz.* 26, pp. 1–34.

— (1945). *Treatise on Thermodynamics (A. Ogg Trans.)* Dover Publications.

Potochnik, Angela (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Prunkl, Carina and Christopher Timpson (n.d.). *Black Hole Entropy is Thermodynamic Entropy*. DOI: https://doi.org/10.48550/arXiv.1903.06276.

Prunkl, Carina E. A. (2020). “On the Equivalence of von Neumann and Thermodynamic Entropy”. In: *Philosophy of Science* 87.2, pp. 262–280. DOI: 10.1086/707565.

Read, James (2020). “Functional Gravitational Energy”. In: *The British Journal for the Philosophy of Science* 71.1, pp. 205–232. DOI: 10.1093/bjps/axx048.

Reichenbach, H. (1956). *The Direction of Time*. University of California Press.

Roach, Ty N.F. (2020). “Use and Abuse of Entropy in Biology: A Case for Caliber”. In: *Entropy* 22.12, p. 1335. DOI: 10.3390/e22121335.

Robertson, Katie (2020). “Asymmetry, Abstraction, and Autonomy: Justifying Coarse-Graining in Statistical Mechanics”. In: *The British Journal for the Philosophy of Science* 71.2, pp. 547–579. DOI: 10.1093/bjps/axy020.

Rosen, Joseph (2008). “Symmetry Rules: How Science and Nature are founded on Symmetry”. In.

Rovelli, C (1993). “Statistical mechanics of gravity and the thermodynamical origin of Time”. In: *Classical and Quantum Gravity* 10.8, pp. 1549–1566. DOI: 10.1088/0264-9381/10/8/015.

Rovelli, Carlo and Matteo Smerlak (2011). “Thermal time and Tolman–Ehrenfest effect: ‘temperature as the speed of time’”. In: *Classical and Quantum Gravity* 28.7, p. 075007. DOI: 10.1088/0264-9381/28/7/075007.

Ruetsche, Laura (2011a). *Interpreting quantum theories*. Oxford University Press.

Ruetsche, Laura (2011b). “Why be normal?” In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42.2. Philosophy of Quantum Field Theory, pp. 107–115. ISSN: 1355-2198. DOI: https://doi.org/10.1016/j.shpsb.2011.02.003. URL: https://www.sciencedirect.com/science/article/pii/S1355219811000153.

Schroeder, D.V. (2021). *An Introduction to Thermal Physics.* Oxford University Press. ISBN: 9780192895547. URL: https://books.google.com/books?id=IRUOEAAAQBAJ.

Schwarzschild, Karl (1916). “On the gravitational field of a mass point according to Einstein’s theory”. In: *Sitzungsber.Preuss.Akad.Wiss.Berlin (Math.Phys.)* Translated by S. Antoni and A. Loinger (1999). Available at https://arxiv.org/abs/physics/9905030., pp. 189–196.

Seidenfeld, Teddy (1986). “Entropy and Uncertainty”. In: *Philosophy of Science* 53.4, pp. 467–491.

Shenker, Orly (1999). “Is $-ktr(\rho log \rho)$ The Entropy in Quantum Mechanics?” In: *British Journal for the Philosophy of Science* 50, pp. 33–48.

Sklar, Lawrence (1993). *Physics and Chance: Philosophical issues in the foundations of statistical mechanics.* Cambridge: Cambridge University Press.

Stewart, Seán M. and R. Barry Johnson (2016). *Blackbody Radiation: A History of Thermal Radiation Computational Aids and Numerical Methods.* CRC Press. DOI: 10.1201/9781315372082. URL: https://doi.org/10.1201/9781315372082.

Stöltzner, Michael (2002). “What Lakatos Could Teach the Mathematical Physicist”. In: *Appraising Lakatos: Mathematics, methodology, and the man.* Ed. by George Kampis, Ladislav Kvasz, and Michael Stöltzner. Dordrecht: Kluwer.

Sutcliffe, W. G. (Sept. 1965). “Lorentz transformations of thermodynamic quantities”. In: *Il Nuovo Cimento (1955-1965)* 39.2, pp. 683–686. DOI: 10.1007/BF02735833. URL: https://doi.org/10.1007/BF02735833.

Swanson, Noel (2014). “Modular Theory and Spacetime Structure in QFT”. Princeton University. PhD thesis.

— (2021). “Can quantum thermodynamics save time?” In: *Philosophy of Science* 88.2, pp. 281–302. DOI: 10.1086/711569.

Swinburne, James (1904). *Entropy, or, Thermodynamics from an Engineer's Standpoint: And the Reversibility of Thermodynamics*. Westminster: Constable.

Takesaki, M. (1970). "Tomita's theory of Modular Hilbert algebras and its applications". In: *Lecture Notes in Mathematics*. DOI: 10.1007/bfb0065832.

Taylor, Henry and Peter Vickers (2017). "Conceptual Fragmentation and the Rise of Eliminativism". In: *European Journal for Philosophy of Science* 7.1, pp. 17–40. DOI: 10.1007/s13194-016-0136-2.

Thebault, Karim P. Y. (2021). "The Problem of Time". In: *The Routledge Companion to Philosophy of Physics*. Ed. by Eleanor Knox and Alastair Wilson. Routledge, pp. 386–400.

Thomson, William (1848). "On an Absolute Thermometric Scale Founded on Carnot's Theory of the Motive Power of Heat, and Calculated from Regnault's Observations." In: In Thomson (1882), 100–106. Originally published in the Proceedings of the Cambridge Philosophical Society 1:66–71; also in the Philosophical Magazine, 3rd ser., 33:313–317.

— (1882). *Mathematical and Physical Papers. Vol. 1.* Cambridge: Cambridge University Press.

Tolman, Richard C. (1934). *Relativity Thermodynamics and Cosmology*. Oxford: Clarendon Press.

Turck-Chièze, S (Jan. 2016). "The Standard Solar Model and beyond". In: *Journal of Physics: Conference Series* 665.1, p. 012078. DOI: 10.1088/1742-6596/665/1/012078. URL: https://dx.doi.org/10.1088/1742-6596/665/1/012078.

Turck-Chièze, Sylvaine and Sébastien Couvidat (Aug. 2011). "Solar neutrinos, helioseismology and the solar internal dynamics". In: *Reports on Progress in Physics* 74.8, p. 086901. DOI: 10.1088/0034-4885/74/8/086901. URL: https://dx.doi.org/10.1088/0034-4885/74/8/086901.

Uffink, Jos (2001). "Bluff Your Way in the Second Law of Thermodynamics". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32.3, pp. 305–394. DOI: 10.1016/s1355-2198(01)00016-8.

Uffink, Jos (2006). “Insuperable Difficulties: Einstein’s Statistical Road to Molecular Physics”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 37.1, pp. 36–70. DOI: 10.1016/j.shpsb.2005.07.004.

Unruh, W. G. (1997). “Time, gravity, and quantum mechanics”. In: *Time’s Arrows Today: Recent Physical and Philosophical Work on the Direction of Time*. Ed. by S. F. Savitt. Cambridge University Press, pp. 23–94.

von Mosengeil, Karl (1907). “Theorie der stationaren Strahlung in einem gleichformig bewegten Hohlraum”. In: *Ann Phys.* 22. Reprinted in Planck (1958), Vol. II, 138–75., pp. 876–904.

Wald, Robert M. (2001). “The Thermodynamics of Black Holes”. In: *Living Reviews in Relativity* 4.6.

Wallace, David (2015). “The quantitative content of statistical mechanics””. In: *Studies in History and Philosophy of Modern Physics* 52, pp. 285–293.

— (2018). “The case for black hole thermodynamics part I: Phenomenological thermodynamics”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 64, pp. 52–67. ISSN: 1355-2198. DOI: https://doi.org/10.1016/j.shpsb.2018.05.002. URL: https://www.sciencedirect.com/science/article/pii/S1355219817301661.

— (2020). “The Necessity of Gibbsian Statistical Mechanics”. In: *Statistical Mechanics and Scientific Explanation*. Chap. Chapter 15, pp. 583–616. DOI: 10.1142/9789811211720_0015. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789811211720_0015. URL: https://www.worldscientific.com/doi/abs/10.1142/9789811211720_0015.

Werndl, Charlotte (2009). “Justifying definitions in mathematics: going beyond Lakatos”. In: *Philosophia Mathematica* 17.3, pp. 313–340.

Wuthrich, C. (2019). “Are Black Holes About Information?” In: *Why trust a theory?: Epistemology of fundamental physics. essay*. Ed. by R. Dardashti, R. Dawid, and K. Thebault. Cambridge University Press.

Wüthrich, Christian (2018). *Are Black Holes About Information?" In: Why Trust A Theory?: Epistemology of Fundamental Physics*. Ed. by Radin Dardashti, Richard Dawid, and Karim Thebault. New York, NY: Cambridge University Press, pp. 202–223.

Zurek, W. H. and Kip S. Thorne (May 1985). "Statistical Mechanical Origin of the Entropy of a Rotating, Charged Black Hole". In: *Phys. Rev. Lett.* 54 (20), pp. 2171–2175. DOI: 10.1103/PhysRevLett.54.2171. URL: https://link.aps.org/doi/10.1103/PhysRevLett.54.2171.