European Journal for Philosophy of Science (2022) 12: 41
https://doi.org/10.1007/s13194-022-00466-w

PAPER IN PHILOSOPHY OF THE NATURAL SCIENCES

Understanding probability and irreversibility in the Mori-Zwanzig projection operator formalism

Michael te Vrugt[1]

Received: 8 December 2021 / Accepted: 30 April 2022 / Published online: 22 June 2022

Abstract

Explaining the emergence of stochastic irreversible macroscopic dynamics from time-reversible deterministic microscopic dynamics is one of the key problems in philosophy of physics. The Mori-Zwanzig (MZ) projection operator formalism, which is one of the most important methods of modern nonequilibrium statistical mechanics, allows for a systematic derivation of irreversible transport equations from reversible microdynamics and thus provides a useful framework for understanding this issue. However, discussions of the MZ formalism in philosophy of physics tend to focus on simple variants rather than on the more sophisticated ones used in modern physical research. In this work, I will close this gap by studying the problems of probability and irreversibility using the example of Grabert's time-dependent projection operator formalism. This allows to better understand how general proposals for understanding probability in statistical mechanics, namely (a) quantum approaches and (b) almost-objective probabilities, can be accomodated in the MZ formalism. Moreover, I will provide a detailed physical analysis, based on the MZ formalism, of various proposals from the philosophical literature, such as (a) Robertson's theory of justifying coarse-graining via autonomous macrodynamics, (b) Myrvold's problem of explaining autonomous macrodynamics, and (c) Wallace's simple dynamical conjecture.

Keywords Statistical physics · Irreversibility · Projection operators · Philosophy of physics · Probability · Mori-Zwanzig formalism

✉ Michael te Vrugt
michael.tevrugt@uni-muenster.de

[1] Institut für Theoretische Physik, Center for Soft Nanoscience, Philosophisches Seminar, Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany

1 Introduction

The philosophy of statistical mechanics is a huge and rich field concerned with a variety of questions. Arguably among the most important of these are:

1. What is the meaning of the probability distributions employed in statistical physics?
2. How does macroscopic irreversibility arise from time-reversal invariant microscopic dynamics?

Both questions are related to the observation that the microscopic equations of motion - both in classical and in quantum mechanics - are deterministic and invariant under time-reversal. Determinism appears to be in conflict with the existence of (objective) probability distributions, while time-reversal invariance appears to be in conflict with the existence of a clear arrow of time. Due to the explanatory role that probability distributions play in statistical mechanics (in particular, the thermodynamic arrow of time is often thought to be related to the initial probability distribution of the universe (Wallace, 2011)), these problems are tied together (Brown, 2017).

Philosophers of physics concerned with irreversibility in statistical mechanics often focus on the history of this field, for example by analyzing the origin and meaning of Boltzmann's H-theorem. While the value of such investigations is undeniable, nonequilibrium statistical mechanics has made and is continue to make considerable progress since Boltzmann's times, and a purely historical analysis is at risk of overlooking important qualitative insights that can be gained from modern theories. In particular, nonequilibrium statistical mechanics is very successful in making quantitative predictions for the approach to equilibrium, which suggests that there is some value in what is done there despite the many "philosophical" objections against the coarse-graining methods employed there (Wallace 2015, 2021).

Recent attempts to close this gap on the philosophical side, in particular from the works of Wallace (2015, 2021) and Robertson (2020), have focused on the *Mori-Zwanzig (MZ) projection operator formalism* (Nakajima, 1958; Mori, 1965; Zwanzig, 1960). The MZ formalism, which is one of the most important coarse-graining techniques used in modern statistical mechanics, allows for the systematic derivation of (irreversible) macroscopic transport equations based on known (reversible) microscopic dynamics. Therefore, a systematic analysis of the way in which this is done allows for an improved understanding of the origin of irreversibility in general. Philosophical discussions of the MZ formalism, however, tend to focus on very simple variants and are therefore still somewhat detached from the way in which it is used in physical research.

In this work, I will analyze the origin of probability and thermodynamic irreversibility in Grabert's time-dependent projection operator formalism (Grabert, 1982), which forms the basis for many applications of the MZ formalism in modern physical research. This allows for a qualitative and quantitative evaluation of various claims made concerning this formalism (and probability and irreversibility in general) in the philosophical literature.

The analysis of the MZ formalism will lead to five main conclusions, two of which are related (mainly) to probability and three of which are related (mainly) to irreversibility. Each of them will be discussed in a separate chapter. These are

- **Probability (a)**: The question whether probabilities in statistical mechanics should be understood in an epistemic or an ontic way can be answered with "both", since the MZ formalism requires two probability distributions (ρ and $\bar{\rho}$). The distribution ρ is the actual (quantum-mechanical) density operator of the system (as suggested by Wallace (2021)), whereas the relevant density $\bar{\rho}$ is constructed on an information-theoretical basis (as suggested by Jaynes (1975a)).
- **Probability (b)**: An alternative interpretation (based on Myrvold (2021)) would interpret $\rho(t)$ as the time evolute of our initial credences and $\bar{\rho}(t)$ as the convenient replacement we use for $\rho(t)$.
- **Irreversibility (a)**: Coarse-graining can be justified not only (as suggested by Robertson (2020)) if it is used to reveal autonomous macro-dynamics, but also if it is used due to limitations of human observers.
- **Irreversibility (b)**: The explanation of autonomous macrodynamics remains a central problem in the physics of irreversibility (Myrvold, 2020). This problem is not identical to the issue of reconciling reversible microdynamics with thermodynamics since it persists if the microdynamics is irreversible.
- **Irreversibility (c):** Wallace's (2011) forward compatibility criterion is assessed quantitatively (showing that it requires simple initial densities and rapidly decaying memory kernels) and qualitatively (showing that simple initial densities may be postulated also at the beginning of experiments).

This article is structured as follows: In Section (2), I will introduce the philosophical debate concerned with probability and irreversibility in statistical mechanics. The MZ formalism is introduced in Section (3). Then, I defend in turn the five conclusions listed above, namely probability (a) in Section (4), probability (b) in Section (5), irreversibility (a) in Section (6), irreversibility (b) in Section (7), and irreversibility (c) in Section (8). I conclude in Section (9).

2 The problem(s) of irreversibility

2.1 Probability

A central issue in the philosophy of statistical mechanics is the understanding of probability. Since it describes systems consisting of many particles whose microscopic state is unknown, statistical mechanics operates with *probability distributions*. In the classical case, the probability distribution $\rho(t)$ is typically taken to describe the probability that the system is, at a time t, at a certain point in phase space. In the quantum case, a system is instead described using a density operator (also known as "density matrix" or "density operator") that in "textbook statistical mechanics" is typically introduced as

$$\hat{\rho} = p_i |\psi\rangle_i \langle\psi|_i, \tag{1}$$

where p_i is the probability that the system is in state described by the wavefunction $|\psi\rangle_i$. Thereby, $\hat{\rho}$ looks a bit like a probability distribution over wavefunctions, and is often also thought of in this way. However, as will be discussed below, whether $\hat{\rho}$ actually is something like a probability distribution is very controversial. From now on, I will drop the hat.

The understanding of probability is a long-stand problem in philosophy (see Hájek (2019) for a detailed review). In the philosophy of statistical mechanics (but not only there), it is common to distinguish between *objectivist* and *subjectivist* approaches to probability (see, e.g., Brown (2017) and Myrvold (2011)). The debate is then concerned with whether probabilities assigned to microscopic states of a many-particle system are objective or subjective. It is then seen as a problem of "subjectivist" approaches that probabilities in statistical mechanics are often taken to have an explanatory role in, e.g., the approach to thermodynamic equilibrium, which is an objective feature of the world (Albert 1994a, b). On the other hand, one might wonder where an "objective probability distribution" might come from given that the microscopic dynamics is deterministic (Brown, 2017).

A part of the problem is that it is not entirely clear what one could mean by an "objective probability distribution" in the context of statistical mechanics. A typical idea would be to define "objective probability" in terms of relative frequencies, but this approach is known to have a variety of problems (La Caze, 2016). Von Kutschera (1969) comes to a similar conclusion regarding the frequentist approach and argues that a proper interpretation of "objective probability" in the natural sciences should include a subjective element. Myrvold (2011) argues that the dichotomy between objective and subjective probabilities does no justice to statistical mechanics and therefore argues that one should use here "almost objective probabilities" (see below) based on human credences, but admits that these by itself will not directly explain any thermodynamic behavior (which is what probability distributions in statistical mechanics are often used for). And Wallace (2021, p. 12) comes to the conclusion that (in classical statistical mechanics) objective probability is "a mysterious concept". It thus appears as if "objective" and "subjective" probability are maybe not the concepts we should distinguish between.

A more precise terminology would be to distinguish (as is common, see, e.g., Frigg (2008)) between *epistemic probabilities* and *ontic probabilities*. Epistemic probabilities represent degrees of belief (these can be *objective*, i.e., the same for every rational observer with the same evidence, or *subjective*, i.e., not fully determined by the evidence), whereas ontic probabilities represent aspects of the physical world (these can be frequencies, propensities, or parts of a Humean best system). When we argue that a probability should be objective in order to be able to play an explanatory role in physics, what we really mean is that it should be a *physical* (ontic) probability since only physical circumstances can explain physical effects[1] A probability assignment based on degrees of belief might be objective in the sense that (almost) every

[1] I leave aside here the question whether there can be non-physical causes of physical effects (such as minds or gods).

rational agent has the same degree of belief in a certain situation, but these degrees of belief still cannot explain any process in the real world. Note that epistemic and ontic concepts of probability do not exclude each other, one can understand both of them as legitimate concepts and simply ask which is the most appropriate one in a specific context (in fact, this is the approach that will be defended in this article). Therefore, philosophers often distinguish between "chances" and "credences" (see Myrvold (2021) for an introduction).

I now present in more detail three ways in which the probabilities used in statistical mechanics can be understood[2]:

1. The probabilities in statistical mechanics arise from quantum-mechanical probabilities (Albert 1994a, b; Wallace 2021).
2. The probabilities in statistical mechanics represent our knowledge about the system (Jaynes 1975a, b).
3. The probabilities in statistical mechanics are "almost objective probabilities" or "epistemic chances" (Myrvold 2012, 2021).

The first option comes in different forms. I will discuss two of them here. First, David Albert (1994a, b, 2000) has suggested that the spontaneous collapses of the wavefunction postulated by Ghirardi-Rimini-Weber (GRW) theory (Ghirardi et al., 1986) could allow to explain thermodynamic irreversibility. In the GRW theory, it is assumed that the wavefunction of a particle is, at a certain rate, multiplied by a Gaussian. This leads to a localization and (effectively) to a collapse. Since GRW theory involves objective stochasticity, one would then have a straightforward explanation for the existence of objective probabilities in statistical mechanics. However, recent computer experiments by te Vrugt et al. (2021b) indicate that Albert's suggestion is not successful as an explanation of thermodynamic irreversibility even if GRW theory is true. Albert (1994a, b) argues that the GRW collapses will bring a system starting in an "abnormal" initial state[3] into a state that leads to a normal thermodynamic time evolution. However, te Vrugt et al. (2021b) have found no such effect in their simulations. Consequently, this approach will not be considered further in this work.

Second, David Wallace (2021) has argued that the probabilities of statistical mechanics arise from "standard" quantum mechanics without spontaneous collapses. His starting point is the observation that the interpretation of the density operator as a "probability distribution over wavefunctions" is not generally possible. If we consider a system consisting of two particles (or, more generally, subsystems) A and B, then A and B will typically be entangled, i.e., the wavefunction describing the state of the joint system cannot be written as a product of a wavefunction for A and a wavefunction for B. In fact, there is no wavefunction that can describe all possible measurements on A, and the correct description of the state of A is a density operator

[2] This is by no means an exhaustive list of the options suggested in the literature. There is, for example, an interesting approach based on incorporating probabilities in a Humean best system (Frigg & Hoefer, 2015; Frigg, 2016). Here, I have chosen three options that fit particularly well to the mathematical formalism discussed in Section (3.2).

[3] An abnormal initial state is a state that, if evolved forward in time using Hamiltonian dynamics, leads to anti-thermodynamic behavior.

obtained by taking the trace over the degrees of freedom of B. This density operator, however, cannot be interpreted as a probability distribution over possible states of A. Consequently, the use of density operators in quantum mechanics is required because of entanglement regardless of any considerations about probability, such that "there is nothing formally novel about their introduction in statistical mechanics" (Wallace, 2021, p. 18). Moreover, a probability distribution over mixed states is mathematically indistinguishable from an individual mixed state. Thus, given that essentially all systems of interest to statistical mechanics are entangled with their environment, we can interpret the mixed states used in (quantum) statistical mechanics as states of *individual systems* rather than as probability distributions over possible pure states.

I now turn to the second option, which I explain following Frigg (2008). According to Jaynes (1975a, b), the probability distributions of statistical mechanics represent our knowledge about a system. Suppose that a random variable x is continuous and can take values in an interval $[a, b]$. Then, the probability distribution $p(x)$ should be chosen in such a way that it maximizes the Shannon entropy

$$S_{\mathrm{S}} = -\int_a^b dx\, p(x) \ln(p(x)) \tag{2}$$

subject to macroscopic constraints of the form

$$\langle f \rangle = \int_a^b dx\, f(x) p(x) = c, \tag{3}$$

which express that (according to our macroscopic evidence) the mean value of an observable f is equal to c. Notably, although this approach is commonly denoted "subjectivist", the probabilities in Jaynes' theory are determined by the available data and do therefore not (solely) represent the personal opinions of individual observers. Thus, "epistemic" is the more appropriate terminology (Frigg, 2008).

The third approach, introduced by Myrvold (2012) and discussed at length in a recent monograph (Myrvold, 2021), is based on the method of arbitrary functions. Here, the idea is that the time evolutes of an (almost) arbitrary initial probability distribution will give the same results for the probabilities of certain (macroscopically feasible) measurements. These probabilities then are "almost objective". As an example, suppose that a gas is initially (at time s) confinend to the left half of a box and then allowed to expand, and that an agent Alice has some credence regarding the state of the gas (represented by a certain probability distribution $\rho_A(s)$). Let $\rho_A(t)$ be the result of evolving $\rho_A(s)$ forward in time using the Liouville equation to a time t (sufficiently long after s to allow for equilibration). Typically, $\rho_A(t)$ will be extremely complicated. The same holds for $\rho_B(t)$, which is the result of evolving forward in time the initial credences $\rho_B(s)$ of an agent Bob who believes that the gas was initially confined to the right of the box. However, $\rho_A(t)$, $\rho_B(t)$, and the equilibrium distribution ρ_{eq} will give the same probabilities for all macroscopic measurements. In fact, these equilibrium probabilities arise from almost every initial credence function (all except for those that would require very detailed knowledge about then microscopic state), making these probabilities almost objective (Myrvold, 2012).

2.2 Coarse-graining

Suppose now that we have found a probability distribution ρ on phase space (in the classical case) or a density operator ρ (in the quantum case) that describes our system. One can then define the *Gibbs entropy*[4] as

$$S = -k_B \operatorname{Tr}(\rho \ln \rho), \tag{4}$$

where k_B is the Boltzmann constant. As is common in statistical mechanics, we use the trace Tr to denote an integral over all phase-space coordinates (known as "classical trace" (Löwen, 1994, p. 253)) in the classical case and a quantum-mechanical trace in the quantum case, allowing to write Eq. (4) in the same form for classical and quantum mechanics. In the quantum case, the entropy defined by Eq. (4) is the *von Neumann entropy*. It is common in *equilibrium* statistical mechanics to use the von Neumann entropy as the statistical-mechanical analogue of the thermodynamic entropy, whereas this is slightly controversial in philosophy (see Hemmo and Shenker (2006) and Shenker (1999) for arguments against and Chua (2021) and Henderson (2003) for arguments in favor of this view). An important property of both the classical and the quantum Gibbs entropy is that it is constant in a system governed by Hamiltonian mechanics (which is invariant under time reversal).

This is problematic since in macroscopic thermodynamics the entropy is *not* constant. It increases and reaches its maximum in the equilibrium state. Thus, a central challenge of Gibbsian statistical mechanics is to make the entropy (4) change (Frigg, 2008). A common way to do this is to replace ρ by an "averaged" density $\bar{\rho}$ that is typically referred to as the *coarse-grained density*. (In contrast, ρ is then called the *fine-grained density*.) If we replace ρ by $\bar{\rho}$ in Eq. (4), we get the *coarse-grained entropy* (as opposed to the *fine-grained entropy* defined in terms of ρ). Unlike the fine-grained one, the coarse-grained entropy can increase. This move is often presented in the philosophical literature in a way that suggests that it were a historical sequence[5] (in the sense that someone first wrongly suggested the fine-grained entropy and that it was later found that one has to coarse-grain). However, coarse-graining was already a part of the original treatment by Gibbs (1902), who was well aware of the problems associated with the constancy of the fine-grained entropy, and a notion of coarse-graining (partitioning of phase space into cells) plays a role also in the even older Boltzmann approach (see Frigg (2008) for a review). The fact that only the coarse-grained entropy increases, e.g., during the expansion of an isolated gas, can be understood also from its role in thermodynamics, where it is related to our ability to extract work from the system - the expansion of the gas would not reduce the ability of an omniscient being with unlimited powers of manipulation to extract work, but it does reduce ours (Myrvold, 2021, p. 161). A considerable part of the

[4] Gibbs, being a frequentist, thought of the probability ρ as measuring the fraction of systems in an ensemble (hypothetical set of infinitely many copies of a system) that are in a particular state (Myrvold, 2016, p. 585). Nowadays, one typically distinguishes between Gibbsian and Boltzmannian statistical mechanics, with the former being based on ensembles. This issue is not essential for this article, see (Myrvold, 2016; Frigg, 2008) for a discussion of this distinction and its relation to the problem of probabilities.

[5] My thanks to an anonymous reviewer for pointing out this problem.

debate in the philosophy of statistical mechanics has been concerned with the question how to justify coarse-graining (although one could reasonable object that *not* coarse-graining is in equal or even stronger need of justification, see Section (6)). A typical justification for the replacement $\rho \to \bar{\rho}$ is that this replacement corresponds to ignoring microscopic details that we cannot measure anyway; a common objection is then that any irreversibility that results from this coarse-graining is an artefact that is therefore illusory and/or anthropocentric (Robertson, 2020).

To understand this situation in more detail, we should take into account that the aforementioned "problem of irreversibility" actually consists of a variety of sub-problems. Following te Vrugt (2021), I distinguish between five "problems of irreversibility":

- **Q1:** What is the location of irreversibility within thermodynamics?
- **Q2:** What is the definition of "equilibrium" and "entropy"?
- **Q3:** What is the justification of coarse-graining?
- **Q4:** Why do systems that are initially in a nonequilibrium state approach equilibrium?
- **Q5:** Why do system approach equilibrium in the future, but not in the past?

Q1 is concerned entirely with the macroscopic theory of phenomenological thermodynamics, and asks which of its axioms actually makes the theory irreversible (see Brown and Uffink (2001), Luczak (2018), and Robertson (forthcoming)). Q2 asks how we should define the entropy, in particular whether we should define it in terms of ρ or in terms of $\bar{\rho}$ (which, if we assume the equilibrium state to be the one with the maximal entropy, also implies different definitions of "equilibrium"). Q3 then asks why it is justified to replace ρ by $\bar{\rho}$. Since this replacement not yet implies an increase of entropy, Q4 then asks why the entropy increases (and not, for example, decreases or remains constant). And since this explanation will, due to the time-reversal symmetry of the underlying microdynamics, often also be applicable to the past, Q5 finally is concerned with why entropy increase only takes place in one direction of time. A common answer to Q5 is the "past hypothesis", in which the thermodynamic asymmetry is explained via an assumption about the initial state of the universe (typically the assumption that the entropy of the early universe was very low) (Albert, 2000). (See Frisch (2005a), Wallace (2011), Brown (2017), and Farr (2021) for a further discussion of the past hypothesis.)

3 Mori-Zwanzig formalism

3.1 The Mori-Zwanzig formalism in philosophy

Having discussed the general theory of coarse-graining, we now come to one of the most important coarse-graining methods used in modern physics, namely the *Mori-Zwanzig (MZ) projection operator formalism*, developed by Mori (1965), Zwanzig (1960), and Nakajima (1958). It has a large number of applications in modern physics, including (but not limited to) active matter (Han et al., 2021), dynamical density functional theory (Español & Löwen, 2009; te Vrugt et al., 2020), general relativity (te Vrugt et al., 2021a), glasses (Das, 2004), high-energy physics

(Huang et al., 2011), and solid-state theory (Fulde, 1995). Introductions to the formalism can be found in Grabert (1982), te Vrugt and Wittkowski (2020a), Rau and Müller (1996), Klippenstein et al. (2021), Schilling (2022), and Zwanzig (2001). Essentially, the MZ formalism allows to describe the dynamics of a many-particle system in terms of the closed subdynamics of an arbitrary set of “relevant variables” $\{A_j\}$. The central idea here is that all variables that can be used to describe the system form a Hilbert space, and the relevant variables form a subspace. (This Hilbert space of observables, which is a convenient mathematical construction, is not to be confused with the Hilbert space of quantum states.) One can now introduce a scalar product and, based on this, a projection operator that allows to project the full dynamics onto the subspace of the relevant variables. The irrelevant part of the dynamics then enters the dynamics via memory and noise terms. As a result, one gets a closed and exact transport equation for the relevant variables. If one approximates the memory term by a memoryless contribution, one gets irreversible dynamics. Consequently, the MZ formalism provides a highly useful tool for studying the microscopic origins of thermodynamic irreversibility (te Vrugt & Wittkowski, 2020a).

This usefulness has not gone unnoticed in the foundations of physics. In his book on the arrow of time in physics, Zeh (2007) has devoted a chapter to the MZ formalism, analyzing in detail how it leads from reversible to irreversible dynamics and which assumptions are involved there. A shorter and less technical discussion was provided by Sklar (1995). Rau and Müller (1996) have provided a detailed review of how the MZ formalism allows to study the emergence of irreversibility. Later, Wallace (2015, 2021) has discussed the MZ formalism as a paradigmatic case of a quantitative method in nonequilibrium statistical mechanics. Finally, Robertson (2020) has used this formalism (which she refers to as “Zwanzig-Zeh-Wallace (ZZW) framework”[6]) as a basis for the position that coarse-graining in statistical mechanics should aim at revealing autonomous macroscopic dynamics.

Wallace (2011) develops the following mathematical understanding of coarse-grained dynamics: In general, a microscopic density ρ can be evolved forwards in time using the microscopic dynamics U. For any coarse-graining procedure C, one can define the $C+$ dynamics as follow: Apply coarse-graining to the microscopic distribution, evolve it for a short time Δt using the microdynamics, coarse-grain again, evolve for Δt again etc. A distribution ρ is said to be forward compatible with C if evolving it using U and then coarse-graining at the end gives the same result as evolving it using the $C+$ dynamics. Hence, an initial density $\rho(s)$ is forward compatible if the diagram

$$\begin{tikzcd}
\rho(s) \arrow[r, "U"] \arrow[d, "P^\dagger"] & \rho(t) \arrow[d, "P^\dagger"] \\
\bar{\rho}(s) \arrow[r, "C+"] & \bar{\rho}(t)
\end{tikzcd}$$

[6] This terminology will not be used here for two reasons. First, the name “Mori-Zwanzig formalism” is *way* more common, in particular in the physics literature. Second, the name “ZZW” framework gives credit not only to authors who developed the formalism (Zwanzig), but also to those who “only” gave an analysis of the formalism in the context of irreversibility (Zeh and Wallace). Thanks to an anomymous reviewer for pointing out these problems.

commutes. (Diagram adapted from Robertson (2020, p. 557).) Wallace (2011) then introduces the *simple dynamical conjecture*, which states that any distribution that is "simple" is forward compatible[7] with C. He does not give a precise definition of simplicity, but suggests that a distribution that can be specified in a closed form as a uniform distribution over certain macroproperties is simple whereas one that can be specified only by time-evolving another distribution[8] is not (Wallace, 2011, p. 19). Based on the simple dynamical conjecture, Wallace then introduces the *simple past hypothesis* which, in the quantum-mechanical form, assumes that the initial quantum state of the universe is simple. This then explains the physical arrow of time. (Note that this initial quantum state is not necessarily pure. Chen (2021), for example, has argued that the universe's quantum state is impure.)

A more specific discussion of the MZ formalism can be found in Wallace (2015, 2021). Here, Wallace introduces it as a prototypical example of coarse-graining in nonequilibrium statistical mechanics and presents the standard derivation of the master equation (which is shown here in our notation). The microscopic density ρ obeys

$$\dot{\rho}(t) = -\mathrm{i}L\rho(t), \tag{5}$$

where L is the Liouvillian (defined as $L = \frac{1}{\hbar}[H, \cdot]$ with the reduced Planck constant $\hbar$, the Hamiltonian H, and the commutator $[\cdot, \cdot]$ in the quantum case and as $L = \mathrm{i}\{H, \cdot\}$ with the Poisson bracket $\{\cdot, \cdot\}$ in the classical case) and the dot denotes a time derivative. One defines a projection operator $P^\dagger$ and an orthogonal projection operator $Q^\dagger$ with the property $P^\dagger\rho = \bar{\rho}$, where $\bar{\rho}$ is the relevant part of the density. (We use, following the notation in Grabert (1982), the dagger to distinguish the projection operator $P^\dagger$, which acts on density operators, from the projection operator P introduced later, which acts on observables. The operators P and $P^\dagger$ are simply each others adjoint, and we can calculate $P^\dagger$ explicitly once we know P (Grabert, 1982, p. 16).) This allows, defining[9] $\delta\rho = \rho - \bar{\rho} = Q^\dagger\rho$, to derive the following exact transport equation for $\bar{\rho}$ (see Wallace (2015, p. 292) and Zeh (2007, p. 62)):

$$\dot{\bar{\rho}}(t) = -P^\dagger\mathrm{i}L\bar{\rho}(t) + \int_0^t du P^\dagger\mathrm{i}Le^{-Q^\dagger\mathrm{i}Lu}Q^\dagger\mathrm{i}L\bar{\rho}(t-u) - P^\dagger\mathrm{i}Le^{-Q^\dagger\mathrm{i}Lt}\delta\rho(0). \tag{6}$$

Setting $\delta\rho(0) = 0$ and assuming that the memory kernel vanishes rapidly (Markovian approximation) gives the time-irreversible approximate transport equation (master equation) (Wallace, 2015, p. 292)

$$\dot{\bar{\rho}}(t) = -P^\dagger\mathrm{i}L\bar{\rho}(t) + \left(\int_0^\infty du P^\dagger\mathrm{i}Le^{-Q^\dagger\mathrm{i}Lu}Q^\dagger\mathrm{i}L\right)\bar{\rho}(t). \tag{7}$$

[7] The original statement (Wallace, 2011, p. 19) uses "forward predictable" (which is a slightly stronger requirement defined in Wallace (2011)) instead of "forward compatible". Here, I follow Robertson (2020), who frames her discussion exclusively in terms of forward compatibility. Since forward predictability implies forward compatibility (and since forward compatibility implies forward predictability for macrodeterministic systems) (Wallace, 2011, p. 16), the simple dynamical conjecture as stated here follows from the original formulation and is equivalent with it for most cases of practical relevance. Note that the mathematical analysis in Section (8) is based on the definition used here.

[8] An example for such a state would be the one shown in Fig. 1b of te Vrugt et al. (2021b).

[9] Most philosophers write ρ_{rel} and ρ_{ir} rather than $\bar{\rho}$ and $\delta\rho$. The latter notation, however, is more common in physics and is in particular used by Grabert (1982) whose method this article is based on.

In particular, Wallace (2015, p. 292) emphasizes the importance of the assumption $\delta\rho(0) = 0$, which is a probabilistic assumption about the initial state of the system. Zeh (2007, p. 61) has compared this way of eliminating the irrelevant degrees of freedom to the way in which one eliminates the "advanced" solutions in the theory of electromagnetic waves. (See Frisch (2005b, 2006) for a discussion of the relation between the thermodynamic and the electromagnetic arrow of time.) Note that the term $P^{\dagger}\mathrm{i}L\bar{\rho}$ often vanishes (Zeh, 2007, p. 62).

Based on these considerations, Robertson (2020) has developed a theory of the justification of coarse-graining in statistical mechanics. She argues (Robertson, 2020, p. 556) that this procedure can be justified in three ways - interventionism (the environment implements the projection $P^{\dagger}$), asymmetric microscopic laws (dynamically ensuring $\rho \rightarrow \bar{\rho}$) and special initial conditions (ensuring that the coarse-grained dynamics gives the correct results for the relevant part of the density). She focuses on the third strategy ("special conditions account"). Typically, Robertson argues, coarse-graining is justified based on measurement imprecisisions (we do not know the exact microstate of a system and therefore have to use an averaged, i.e., coarse-grained, description). Based on this, it is frequently objected that the irreversible laws obtained by coarse-graining are anthropocentric and/or illusory. However, Robertson continues, we do in fact not coarse-grain because of measurement imprecisions but because we want to reveal autonomous higher-level macrodynamics. This is what the MZ formalism does, and the projection operator $P^{\dagger}$ has to be constructed in such a way that it leads to such autonomous dynamics. Consequently, the irreversibility of the higher-level transport equation (7) is not illusory, but (weakly) emergent.

3.2 Grabert's projection operator formalism

Wallace and Robertson, while in principle acknowledging the broad applicability of the MZ formalism, have in practice only considered one rather simple variant, namely the derivation of master equations using time-independent projection operators. Hence, one is not explicitly concerned here with individual observables (foundational discussions usually don't even mention the Hilbert-space understanding of the MZ formalism). While the master equation approach is quite general (Grabert, 1982), it is also highly abstract, which has the consequence that studying only this variant leads one to overlooking important issues. Current research on this topic in physics instead focuses on studying the dynamics of individual observables using time-dependent projection operators. These variants of the formalism have been pioneered by Robertson (1966), Kawasaki and Gunton (1973) and Grabert (1978). More recently, extensions have been derived by Meyer et al. (2017, 2019) and te Vrugt and Wittkowski (2019). Here, I will explain the time-dependent projection operator formalism as it is described in the textbook by Grabert (1982), which forms the basis of most of the work that is done today.

The microscopic description of a many-particle system is given by its density operator ρ. Since it is not known exactly, it has to be approximated. For this purpose, one introduces a *relevant density* $\bar{\rho}$. The relevant density is often assumed to have the form (Grabert, 1978, p. 482)

$$\bar{\rho}(t) = \frac{1}{Z(t)} e^{-a_i^\natural(t) A_i} \tag{8}$$

where the partition function $Z(t)$ is a normalization constant (ensuring $\mathrm{Tr}(\bar{\rho}(t)) = 1$) and the thermodynamic conjugates $a_i^\natural(t)$ (this notation is adapted from Wittkowski et al. (2012, 2013)) ensure that the macroequivalence condition $a_i(t) = \mathrm{Tr}(\bar{\rho}(t) A_i)$ holds. (Summation over indices appearing twice is assumed throughout this article.) Here,

$$a_i(t) = \mathrm{Tr}(\rho(t) A_i) \tag{9}$$

is the average of the observable A_i. The form (8) can be justified from an information-theoretical point of view as it expresses maximal noncommittance regarding microscopic details, i.e., it assigns all microscopic configurations that are compatible with the set of macroscopic values $\{a_i(t)\}$.

One can then define the time-dependent projection operator P acting on an arbitrary observable X as (Grabert, 1978, p. 487)

$$P(t)X = \mathrm{Tr}(\bar{\rho}(t)X) + (A_i - a_i(t))\,\mathrm{Tr}\left(\frac{\partial \bar{\rho}(t)}{\partial a_i(t)} X\right). \tag{10}$$

Equation (10) implies (Grabert, 1982, p. 16)

$$P^\dagger(t)\rho(t) = \bar{\rho}(t). \tag{11}$$

As shown in the Appendix, the mean values a_i obey the exact differential equation

$$\dot{a}_i(t) = v_i(t) + \int_s^t du R_{ij}(t,u) a_j^\natural(u) + f_i(t,s) \tag{12}$$

with the (reversible) organized drift v_i, the retardation matrix R_{ij}, and the mean random force f_i. We now make two important assumptions:

1. Markovian approximation: The relevant variables evolve slowly compared to the microscopic degrees of freedom. This implies that the memory kernel in Eq. (12) falls off on a very short timescale, and that the thermodynamic conjugates $a_i^\natural(t)$ are approximately constant on this timescale.
2. The density operator at the initial time $t = s$ is of the relevant form, i.e., $\delta\rho(s) = 0$. This allows to set $f_i = 0$ (as can be seen from the definition of f_i given by Eq. (28) in the Appendix).

This allows to replace Eq. (12) by the approximate transport equation

$$\dot{a}_i(t) = v_i(t) + D_{ij}(t) \frac{\partial S}{\partial a_j(t)} \tag{13}$$

with the diffusion tensor

$$D_{ij}(t) = \frac{1}{k_B} \int_0^\infty du \int_0^1 d\alpha\, \mathrm{Tr}(\bar{\rho}(t) e^{\alpha a_k^\natural(t) A_k} (e^{\mathrm{i}Lu} Q(t) \dot{A}_i) e^{-\alpha a_k^\natural(t) A_k} \dot{A}_j). \tag{14}$$

Formally, the Markovian approximation corresopnds to disregarding terms of third or higher order in $\dot{A}_i$ (which is what allows us to replace $a_j^\natural(u)$ by $a_j^\natural(t)$) (Grabert, 1982, p. 39). We have used that

$$a_j^\natural(t) = \frac{1}{k_B}\frac{\partial S}{\partial a_j(t)} \tag{15}$$

with the (coarse-grained) entropy (Grabert, 1978, p. 483)

$$S = -k_B \,\mathrm{Tr}(\bar{\rho} \ln \bar{\rho}) = k_B \ln Z + k_B a_j^\natural a_j. \tag{16}$$

One can show that

1. the organized drift term v_i does not contribute to the rate of change of the entropy (Grabert, 1978, pp. 483-484).
2. the tensor D_{ij} is, due to the Wiener-Khintchine theorem, positive definite (Español & Löwen, 2009; Anero et al., 2013).

This implies that

$$\dot{S} = k_B^2 D_{ij} a_i^\natural a_j^\natural \geq 0, \tag{17}$$

which shows that the approximate transport equation (13) is irreversible (Anero et al., 2013).

4 Probability (a): the quantum approach

I will now explain what the MZ formalism can contribute to solving the problems of irreversibility and probability in statistical mechanics. On the one hand, I will thereby provide a conceptual understanding of the mathematical formalism outlined in Section (3.2). Moreover, I will use the equations from Section (3.2) in order to give the general considerations from Wallace and Robertson a quantitative underpinning.

First, I will discuss how, within the MZ formalism, we can address the problem of understanding probability in statistical mechanics as introduced in Section (2.1). As will be shown, there are (at least) two options, the first one being based on the idea that ρ is just the quantum-mechanical density operator (as suggested, e.g., by Wallace (2021)), and the second one being based on the idea that probabilities in statistical mechanics are almost-objective probabilities (Myrvold, 2021). In this section, I will consider the quantum approach, which is first discussed in a general way. Then, I will show that incorporating it into Grabert's MZ formalism shows that it still requires information-theoretical elements in the spirit of Jaynes' theory.

The key point to take into account here is that there are *two* densities, not one. First, there is the microscopic density ρ. It describes (at least according to "textbook understanding") the actual state of the system. Second, there is the relevant density $\bar{\rho}$ (typically given by Eq. (8)). It describes our knowledge about the state of the system. Although the existence of these two different distributions is never questioned in the physics literature, it is actually extremely surprising from a classical point of view. Both ρ and $\bar{\rho}$ are probability distributions. If it is the point of probability distributions in statistical mechanics to express ignorance of the system, then what does "correct microscopic probability distribution" even mean? And if this is not the point of probabilities in statistical mechanics, then what is?

In Section (3.2), I have introduced three interpretations of probability in statistical mechanics. Let us start with the first one by taking ρ to be simply the actual density operator of the system. There are two important objections against this view. First, one could ask whether using the density operator really allows us to get probabilities. This has been questioned by Brown (2017, p. 38), who argued that, within the Everettian interpretation of quantum mechanics that Wallace defends, "a density operator (...) is no more intrinsically a carrier of probability than is the Liouville measure on the classical phase space." Probabilities, he argues, arise - in an Everettian framework - only when a rational agent bets on measurement outcomes[10]. This implies that (for an Everettian) "quantum probabilities make no appearance at the start of the world, but are forced on us at the later times at which observations are made" (Brown, 2017, p. 38), and thus seems to suggest that quantum probabilities do not give us statistical probabilities at the start of the world. While Brown's observation is correct, it does not pose any problem for Wallace's approach (at least not when it is applied to the MZ formalism). The only reason we require an interpretation of ρ as (something like) a probability distribution is that we want to interpret the expression $\mathrm{Tr}(\rho X)$ as the mean value of the observable X. This expression is nothing else than the expectation value of a quantum-mechanical measurement of the observable X on a system in the state ρ. Consequently, the probabilities required here are just quantum-mechanical probabilities[11], and if one assumes that the Everett interpretation can explain probabilities, then it is also able to explain the probabilities in statistical mechanics. In particular, if no measurements are made, there is no need for probabilities since then there is nothing probabilistic about the formalism introduced in Section (3.2) - we are simply solving differential equations and making approximations for them. Strictly speaking, the question we should be asking when confronted with the formalism presented in Section (3.2) is not "What is the meaning of the probability distribution in statistical mechanics?", but simply "What is the meaning of the symbol ρ?" This can be answered with "the density operator" regardless of whether or not we take the density operator to represent probabilities. (Note that the we do not need to adapt the Everett interpretation here. *Any* interpretation of quantum mechanics that allows to interpret $\mathrm{Tr}(\rho X)$ as the expectation value of a measurement of X on a system in state ρ - in other words: any interpretation in which the Born rule holds - allows for such an understanding of ρ. Just pick whatever is your favourite interpretation of quantum mechanics.)

Second, one could ask what this implies for classical statistical mechanics, which the MZ formalism is also applied to. While in a real physical system one could argue

[10] It is common in the Everett interpretation to assume that quantum-mechanical probabilities are related to the betting behavior of rational agents (Wallace, 2012).

[11] As Wallace (2021, p. 25) puts it: "there *is* something probabilistic about ρ, and about the forward-compatibility requirement, but only in the sense that there is something probabilistic about the quantum state itself (however that probabilistic nature is to be understood)".

that it is always ultimately described by quantum mechanics[12], thermodynamic irreversibility is also observed in classical molecular dynamics simulations (Tóth, 2022) that involve no quantum effects of any sort. Moreover, one can apply the MZ formalism (like statistical mechanics in general) also to astrophysical (te Vrugt et al., 2021a) or colloidal (Español & Löwen, 2009) systems, and it is not very plausible that the dynamics of macroscopic colloids or even stars depends on quantum effects. Finally, one could argue that, if $\vec{\Gamma}_0$ is the actual microscopic value of the phase-space variables contained in a vector $\vec{\Gamma}$, the microscopic density is simply proportional to $\delta(\vec{\Gamma} - \vec{\Gamma}_0)$ with the Dirac delta distribution δ. This is actually common practice in classical many-body physics (te Vrugt & Wittkowski, 2020b). However, a density given by a Delta distribution will typically not take the form (8), and the assumption that the initial density has this form was quite essential for the derivation of Eq. (13).

To understand this issue, we should clarify what we mean by the mean value a_i given by Eq. (9). It is an ensemble average, and the approximate transport equation (13) describes the dynamics of the ensemble average of the observable A_i. The ensemble average, and the ensemble average only, is monotonously approaching equilibrium. In contrast, the actual value in an individual classical system will typically approach a state corresponding to the macrostate with the largest phase-space volume (Boltzmannian equilibrium), but will continue to fluctuate around this equilibrium state. A good way to see this is to consider the example of *dynamical density functional theory* (DDFT) (te Vrugt et al., 2020), a theory for the one-body density[13] of a classical fluid that exists in deterministic and stochastic forms. The deterministic theory, which can be derived as a special case of Eq. (13) (Español & Löwen, 2009) describes the ensemble-average of the one-body density, its stochastic counterpart describes actual physical systems (Archer & Rauscher, 2004; te Vrugt et al., 2020; te Vrugt, 2021). In deterministic DDFT, equilibrium is approached monotonously, whereas there are fluctuations around equilibrium in stochastic DDFT. If we take the microscopic distribution ρ to be proportional to $\delta(\vec{\Gamma} - \vec{\Gamma}_0)$ in the classical case (such that the ensemble average of an observable is always just its actual value in one specific system), then f_i will never vanish and the dynamics will never, strictly speaking, be forward predictable by the coarse-grained dynamics (which in practice typically implies that it fluctuates away from equilibrium). We can, of course, introduce a “smoother” microscopic distribution by hand, for example by considering a probability distribution over initial conditions (this is what is typically done in classical statistical mechanics). In a computer experiment, one can repeat a classical simulation several times with random initial conditions and calculate the average of an observable over all these simulations. However, the probability distribution this average is taken with respect to might has quite a different interpretation than the quantum density operator ρ, since it is not a physical property of an actual system and can therefore not be used to explain an actual system’s behavior.

[12] Wallace (2021) argues, in particular, that the classical phase-space distribution function arises as a limit of the Wigner function, which is equivalent to the density operator.

[13] By “one-body density”, I mean the number of particles at a certain position, not the microscopic probability density discussed in the rest of this article.

If we take ρ to be the density operator, we are still left with the question what $\bar{\rho}$ is. In the presentation by Wallace (2021), $\bar{\rho}$ is simply what one gets if one applies the projection operator P to ρ. Formally, this is absolutely correct. However, it does not mention an important point about why one constructs the relevant density and the projection operator in the way one does. To see why one has to coarse-grain in this particular way, we have to consider why Eq. (8) has the form it has. The reason is, as discussed in Section (3.2), information theory. The relevant density is, in the spirit of Jaynes, constructed by maximizing the informational entropy. Consequently, while ρ is an ontic probability (or better: state), $\bar{\rho}$ is an epistemic probability distribution even in the quantum case. The question whether probabilities in statistical mechanics are epistemic or ontic thus has a surprising answer: both.

5 Probability (b): Almost-objective probabilities

Thus, we have found an account of probability of that combines the first and second interpretation suggested in Section (2.1), by taking ρ to be an ontic (quantum-mechanical) and $\bar{\rho}$ to be an epistemic (information-theoretical) probability. This account will be referred to as *option (a)*. In this section, we will consider an alternative account, from now on referred to as *option (b)*, which can be constructed based on Myrvold's view that the probabilities in statistical mechanics are almost-objective probabilities. In particular, I will analyze how well Grabert's and Myrvold's approaches fit together despite the fact that the former appears to be more restrictive about the choice of $\bar{\rho}$. As will be shown, this is only an apparent contradiction.

In Myrvold's theory, the initial probability distribution represents one's initial credences about the system. The credences at later times will not be the Hamiltonian time evolutes of the initial one, but instead will be simpler distributions determined by the macrostate of the system. Thus, Myrvold's theory also involves - for a given observer - two probability distributions, namely the Hamiltonian time evolute of the initial credences and the simpler distribution used at later times. We may identify these two distributions with the two distributions appearing in the MZ formalism by assuming $\rho(t)$ to be the Hamiltonian evolute of the initial credences at time t - in line with the fact that, in Section (3.2), we have assumed that ρ evolves according to Eq. (5) - and $\bar{\rho}(t)$ with the actual credences our observer has at time t. In particular, this forces us to set $\rho(s) = \bar{\rho}(s)$ (since our observer has only one credence function at the initial time). Next, we observe that ρ and $\bar{\rho}$ evolve differently. In Eq. (12), the organized drift term v_i is the part of the dynamics that we would have if ρ evolved like $\bar{\rho}$ at all subsequent times, whereas the memory terms (that lead to dissipation and thus equilibration) result from deviations of ρ and $\bar{\rho}$ (Grabert, 1978). Consequently, in a Myrvoldian framework, we can indeed interpret $\rho(t)$ as the time evolute of our initial credences, and $\bar{\rho}$ as the simpler distribution that we use as a surrogate for ρ. This interpretation is very different from the one suggested in Grabert (1982), where ρ is the observer-independent microscopic distribution. In fact, it would here be assumed that $\bar{\rho}$ (and not ρ) is "almost objective" in the sense defined in Section (2.1). However, this interpretation is also compatible with the derivation presented in Section (3.2).

Since Myrvold's theory is not based on Grabert's projection operator formalism, it is an interesting question whether option (b) is consistent with the position that Myrvold advocates. A reason why it might appear as if it was not is that one might have the impression that (1) we have to set $\rho(s) = \bar{\rho}(s)$ in option (b) and (2) $\bar{\rho}$ is always given by Eq. (8). From (1) and (2) it would follow that initial credence functions would always have to be given by Eq. (8), which is in contradiction with Myrvold's approach that allows for much more flexibility in the choice of the credence function. After all, Myrvold's approach is based on the method of *arbitrary* functions, and it is quite essential that any initial credence function will lead to a similar stationary (equilibrium) state as long as it is reasonable. "Reasonable" here means that if can arise from macroscopic measurements and does not require us to postulate detailed knowledge about microscopic correlations like those required for generating anti-thermodynamic time evolutions. Surely this initial credence does not have to be given by Eq. (8).

However, there is in fact no contradiction between Myrvold's position and option (b) proposed here. The reason is that Grabert's approach is very flexible in the choice of $\bar{\rho}$ - it can be any function of the macroscopic variables that satisfies the macroequivalence condition. Therefore, while (1) holds, (2) is wrong. In fact, Grabert (1982, p. 20) even argues that the condition $\rho(s) = \bar{\rho}(s)$ "should be looked upon as a condition for an adequate definition of the relevant probability density rather than a restriction of initial states". The reasons that the form (8) is typically used (and will also be used for the rest of this work) are that it is a natural choice for systems starting in constrained equilibrium (see Section (8)) and that it has technical advantages. For example, it allows (as shown explicitly in the Appendix) to write the general transport equation in the convenient form (12). But this is just a question of mathematical convenience. Apart from this, $\bar{\rho}$ can be chosen according to one's initial credences whatever they might be, as long as they are not "unreasonable" in the sense that they would require very detailed microscopic knowledge. However, a distribution whose specification would require very detailed microscopic knowledge presumably cannot be written as a function of the macroscopic variables anyway, and is therefore not a suitable candidate for $\bar{\rho}$. Note that this point also shows why it is useful to use Grabert's more sophisticated variant of the MZ formalism rather than a simpler one, since simpler ones do not typically allow for such a detailed discussion of the forms that $\bar{\rho}$ can possibly take. Such a detailed discussion, however, is helpful in order to see that and how we can incorporate Myrvold's theory into the general picture of nonequilibrium statistical mechanics painted by the MZ formalism.

The problem with option (b) is that it is assumed in the MZ formalism that, for *any* observable X, the expectation value is given by $\mathrm{Tr}(\rho X)$ - not just for the macroscopic observables, for which $\bar{\rho}$ gives the correct expectation value. This is not guaranteed if X is in any way related to the credences of a human observer. In contrast, in quantum mechanics, one can simply take ρ to be the density operator of the system of interest, which in general will be a mixed state. Moreover, only option (a) allows us to use the initial form of ρ as an explanation for the physical behavior of a system. (A response to this objection against credence-based approaches can be found in Myrvold (2021, pp. 193–197).)

There are, however, also good arguments for option (b). In his monograph on the topic, Myrvold (2021, pp. 224-225) discusses also Wallace's quantum approach to probabilities and acknowledges it as a possible solution to the problem of probability. This leads him to the question "Can classical statistical mechanics stand on its own two feet?". He argues that this is the case, since the dynamics of equilibration in classical statistical mechanics simply requires some uncertainty in the initial conditions, which can come from quantum mechanics or solely from epistemic considerations. These considerations are interesting, first of all, because they point to a unification of options (a) and (b) - or, more generally, simply to the fact that it may depend on the context which of these interpretations is more appropriate. The contribution of the present work is then to point out that and how precisely both options can be embedded in the MZ framework. Option (b) has, in particular, an advantage in addressing the problem of classical molecular dynamics simulations discussed in Section (4) - where, unlike in a real system, there is no quantum mechanics that the classical dynamics is a limit of. Here, expectation values as described by Eq. (9) are usually calculated by repeating a simulation several times with varying initial conditions (Orlandini et al., 2011). The distribution of initial conditions then generates the uncertainty that, in a real system, might ccome from quantum mechanics. This issue will be discussed further in in Section (8).

6 Irreversibility (a): Coarse-graining

We now shift our focus from probability to irreversibility. To keep the argumentation focused, I follow the five-problems scheme by te Vrugt (2021) introduced in Section (2.2) and start with Q3 (the justification of coarse-graining). One can object here, quite reasonably, that "How can we justify coarse-graining?" is not the right way of posing this problem, given that it suggests that using the fine-grained entropy would not be in need of justification. Of course, any choice needs justification, and using the fine-grained entropy perhaps even more since it has properties (constancy in closed systems) that the thermodynamic entropy out of equilibrium is not generally considered to have. However, since coarse-graining has received some very harsh criticism in the philosophical literature – Redhead (1995, p. 31) even called it a "disreputable procedure" – and since previous work which this article is based on (in particular Robertson (2020)) has also framed the topic in this way, it is helpful to stick to this way of formulating the third problem. Robertson (2020, p. 561) has proposed that this issue can be split into two sub-problems - namely, the justification of a *particular* coarse-graining projection and the justification of coarse-graining *in general*.

We start by asking what Grabert's variant of the MZ formalism can, as opposed to the simpler one presented in Section (3.1), tell us about this problem. A first thing to note here is that the derivation presented in Section (3.2) works in the Heisenberg picture, as opposed to the Schrödinger picture derivation of the master equation usually considered in the philosophical literature. This has one key advantage, namely that we are not considering the equations of motion for an abstract density operator, but those for a clearly identifiable set of relevant variables $\{A_i\}$. (Note, however, that the Heisenberg picture is also used in more basic variants of the MZ formalism (te Vrugt & Wittkowski, 2020a).) Thus, we can specify Robertson's two subproblems as follows: Why can we restrict ourselves to a subset $\{A_i\}$ of the system's dynamical variables, and how can we figure out which subset to choose?

As discussed in Section (3.1), Robertson (2020) takes the revelation of autonomous macro-dynamics to be the justification for coarse-graining in general. A particular coarse-graining method then should be chosen in such a way that the system obeys an autonomous dynamics on the macrolevel. She explicitly admits that "this criterion will not help physicists discover new, useful maps", and that the resulting projection operators "will not look especially unified" (Robertson, 2020, p. 568). The latter observation is fully correct for the MZ formalism in general, since different variants use projection operators with very different properties (this even holds if we only look at variants with time-dependent projection operators). However, if we restrict ourselves to Grabert's approach, some degree of unification can be achieved since the relevant density and the projection operator can in general be constructed using Eqs. (8) and (10), respectively. This is an interesting observation since these equations are based on Jaynes' information-theoretic approach. (And this is already quite a general result, since Grabert's formalism can be shown to incorporate several popular variants of the MZ formalism such as Mori theory or the master equation approach (Grabert, 1982).)

The question is then what counts as "autonomous macro-dynamics". By "autonomous dynamics", Robertson (2020, p. 553) means that the dynamics of $\bar{\rho}$ depends neither on $\delta\rho$ nor (explicitly) on t. The explicit time dependence is eliminated by the Markovian approximation, the $\delta\rho$ dependence by the assumption $\delta\rho(s) = 0$. This definition of "autonomous", which is the one used in the theory of dynamical systems, is somewhat unfortunate in this context as it would (combined with the idea that the MZ formalism ought to reveal autonomous dynamics) automatically render all applications of the MZ formalism to systems driven by explicitly time-dependent external potentials (te Vrugt & Wittkowski, 2019) unjustified. Presumably, however, we can understand Robertson's criterion as implying that we should choose the $\{A_i\}$ in such a way that their mean values obey an equation of the form (13), which requires that the Markovian approximation is justified. This requires that the relevant variables are slow compared to the microscopic degrees of freedom. A set of slow variables can typically be constructed by considering both the conserved variables and the variables associated with a spontaneously broken symmetry. As an example, consider the dynamics of a crystal. Here, the slow variables are density and momentum (conserved variables), and the symmetry-restoring low-frequency Goldstone modes of the crystal. Consequently, these are an appropriate set of relevant variables for deriving a theory for the elastic properties of a crystal via the MZ formalism (Walz & Fuchs, 2010; Ras et al., 2020; Haussmann, 2022). This shows that Robertson's criterion is not only useful for practitioners of physics, it is actually already used by them.

Moreover, Robertson argues that revealing autonomous macrodynamics is *the* justification for coarse-graining, while justifications like measurement imprecision are inappropriate. The reason is, Robertson (2020) argues, that one would otherwise face the problem that irreversibility (an effect associated with coarse-graining) would be illusory and/or anthropocentric. As she puts it (Robertson, 2020, p. 565), it "seems unlikely that advances in the science of microscopy will lead to different choices of" (the projection operator). This view is related to the fact that philosophers tend to study coarse-graining almost exclusively in its relation to irreversibility, thereby ignoring its much wider use in physics.

A good example for an application of the MZ formalism that is not related to autonomous macrodynamics but to the limitations of human observers is the study of turbulent fluids. Their dynamics is characterized by a coupling of all length scales (i.e., all wavenumbers) in the system. Small length scales influence the large ones and vice versa. When studying and simulating such a fluid - a problem of great importance in engineering - one faces the problem that simulations can only resolve a finite length scale. This finite resolution then leads to inaccuracies in the simulation results also on large scales.

This problem is frequently addressed using so-called "large eddy simulations" (see Sagaut (2006) for an introduction). In a large eddy simulation, one simulates the large scales explicitly and includes the smaller scales via a subgrid model. Here, the MZ formalism can play a useful role (Parish & Duraisamy, 2017; Maeyama & Watanabe, 2020). One uses the Navier-Stokes equation (which describes the dynamics of incompressible fluids) as a microscopic model and then projects onto the small wavenumbers (i.e., the large length scales), which are the relevant variables. The memory terms then incorporate the small-scale effects.

Notably, this is not done because the large length scales in the fluid obey an autonomous macrodynamics. In fact, it is precisely the problem that they do not. Instead, we use the MZ formalism because our computers are not good enough to solve the full Navier-Stokes equation numerically, i.e., because of limited available microscopic information. Of course, this coarse-graining induces artefacts, which can be seen as anthropocentric. This, however, is unavoidable (although one of course wishes to minimize it). And if "better microscopes" (i.e., better computers) would be available, windpark engineers would certainly use them rather than the less accurate large eddy simulation models.

Similar ideas are relevant for general relativity. Since the Einstein field equations are highly nonlinear and therefore do not commute with an averaging procedure, it is not possible to get an averaged large-scale model of the universe by simply inserting the averaged matter distribution into these equations. However, this is precisely what is done in the derivation of the Friedmann equations. This issue, known as the "averaging problem", is not fully understood and has even been suggested as an explanation for dark energy (Clarkson et al., 2011). Recently, te Vrugt et al. (2021a) have addressed this problem by extending the MZ formalism to general relativity, which allowed them to derive a correction term for the Friedmann equations. Similar to the case of turbulence, this study is motivated not by the existence of autonomous macrodynamics but by the impossibility of actually solving the Einstein field equations for the complicated matter distribution of the universe.

At this point, the following objection could be raised: If the dynamics is not autonomous at all, e.g., if the dynamics of the long wavelengths in turbulent flow strongly depends on that of the small ones and there is no way of deriving a closed subdynamics for them, then it would be pointless to use the MZ formalism for deriving a dynamic equation for the long wavelengths only. The fact that such equations nevertheless are derived and are of some use seems to suggest that there is at least some autonomy. To use a very drastic example: If we drop a sheet of paper from a building during a snowstorm, we clearly cannot keep track of the effects the wind has on its dynamics, but we also cannot just ignore it, so there seems to be little hope for deriving a closed equation of motion that describes solely the paper.[14]

To understand this issue better, it is important to carefully distinguish between *projecting out* the irrelevant degrees of freedom and *ignoring* them. The irrelevant degrees of freedom can, even if they are projected out, have considerable effects on the dynamics as they are what leads to irreversibility (such that, strictly speaking, it is not really appropriate to call them "irrelevant"). As discussed above, they manifest themselves in two ways, namely (1) in the memory term and (2) in the random force. "Autonomous macrodynamics", is, if we use a very strict definition, present if (1) the memory can be removed using the Markovian approximation and (2) the random force can be dropped. So, on a very strict reading of Robertson's view, these are the two conditions that make it reasonable to use the MZ formalism. Let us start with (1). The Markovian approximation essentially corresponds to the assumption that the expression under the memory integral is well approximated by a Dirac delta distribution. Thereby, autonomous equations (in the mathematical sense) for the relevant variables can be derived. For turbulence, however, this is not the case. Parish and Duraisamy (2017) consider a variety of *non-Markovian* closures for the memory kernel, with the simplest one being the t-model (Chorin et al., 2002). Here, the integral from 0 to t in Eq. (12) is replaced by a term that is simply proportional to t. This model already turns out to be quite useful (Chandy & Frankel, 2010).

Let us now turn to (2). The sheet of paper in the snowstorm bears a certain similarity to something that actually is an important application of the MZ formalism, namely the dynamics of a Brownian particle in a fluid. Such a particle is subject to a large number of collisions with fluid particles, making its precise trajectory impossible to predict. Nevertheless, we can make certain statements about what these particles do, in particular about their statistics. As shown by Zwanzig (1973), the MZ formalism allows to derive the governing equation for the momentum $\vec{p}(t)$ of the particle, the so-called *Langevin equation*, which is given by

$$\dot{\vec{p}}(t) = -\gamma \vec{p}(t) + \vec{\xi}(t), \tag{18}$$

where γ is a friction coefficient and $\vec{\xi}$ is delta-correlated noise (originating from the random force term). Evidently, Eq. (18) is not autonomous since the right-hand side contains the explicitly time-dependent function $\vec{\xi}$. Nevertheless, Eq. (18) is one of the most important equations in theoretical soft matter physics – it has even received some coverage in the philosophical literature (Wallace, 2021; Luczak, 2016; Myrvold, 2021) – and constitutes one of the most important applications of the MZ formalism. In fact, exact transport equations obtained via the MZ formalism are often

[14] My thanks to an anonymous reviewer for pointing out this problem and suggesting this example.

referred to as "generalized Langevin equations" (Zwanzig, 1973; Meyer et al., 2017). If this application of the MZ formalism is not justified, then no application is.

Nevertheless, the idea behind the snowstorm objection of course remains correct. In the cases of turbulence or Brownian motion, the memory kernel or the random force cannot be neglected, but they still have relatively simple forms. In general, both can be extremely complicated time-dependent functions, which can be difficult to figure out and which make Eq. (12) impossible to solve. Thus, a refined version of Robertson's approach - which classifies dynamics as "somewhat autonomous" in a yet to be defined sense if the memory kernel is not a Delta distribution, but still a simple function - can probably deal also with turbulence or Brownian motion. Nevertheless, it needs to be a refined version since the macroscopic dynamics is, for both these applications, far from being autonomous. Moreover, the *motivation* for applying the MZ formalism, at least in the case of turbulence, is not this simplicity (this is only the reason that it works), the motivation is still the fact that we cannot resolve small length scales even though we want to.

This does not mean that Robertson's justifying of coarse-graining in the case of irreversible transport equations in nonequilibrium statistical mechanics is not correct - it is fully appropriate for the analysis of irreversibility. The point I wish to make here is that the justification of coarse-graining depends heavily on the context in which it is used, and that "measurement imprecision" is not an illegitimate one - it is necessary to use procedures of this form for designing airplanes or windparks. Studying the justification of coarse-graining case by case is important as it has implications for our understanding of the effects that result from it. For example, the predictions that MZ-based models with simple approximations for the memory kernel make for transfer spectrum in turbulent flows can differ from the actual spectrum (Parish & Duraisamy, 2017, p. 17). This is, like irreversibility, an effect that arises only after coarse-graining. However, in the case of irreversibility, we can - as shown by Robertson (2020, pp. 573–576) - consider it to be not illusory, but (weakly) emergent. It is a consequence of robust autonomous macrodynamics. In contrast, the artefacts in the MZ-based large eddy simulations are not emergent, but simply an (unavoidable) technical error. This is due to the fact that we coarse-grain here not to reveal autonomous macrodynamics, but simply because our human and computational limitations leave us with no other option.

To summarize: We coarse-grain because we wish to study the subdynamics of a certain set of variables in a system we cannot (or do not want to) describe completely. This can be done because of human or technical limitations - as in the case of turbulence - or because we wish to reveal or study autonomous macro-dynamics - as in the case of irreversible statistical mechanics.

7 Irreversibility (b): Approach to equilibrium

I have thus argued that coarse-graining in statistical mechanics has an information-theoretic basis. It is based on what we know about the system or what we are interested in, and it changes if we know more or if we are interested in more. This raises a question: Robertson (2020) introduces "the possibility of revealing autonomous macro-dynamics" as the justification for a particular form of coarse-graining to avoid the problem that, if irreversible equations of motion are found by ignorance-based coarse-graining, irreversibility might be an illusion. Given that I now propose that coarse-graining is based on information theory and that $\bar{\rho}$ represents an epistemic probability distribution, does that imply that I have precisely this problem? The answer is no, and the reason is that Robertson, while not fully answering the third question, is correct about the fourth one. While the set of relevant variables $\{A_i\}$ can in principle be chosen in an arbitrary way, not every such set will be found to obey a closed macroscopic dynamics.

Again, we start by asking ourselves which novel aspects we can learn from Grabert's MZ formalism as opposed to simpler ones. The answer is related to the relevant density $\bar{\rho}$. As discussed in Section (3.2), the Markovian approximation essentially corresponds to a Taylor expansion up to second order in $\dot{A}_i$ which allows to replace $a_j^\natural(u)$ by $a_j^\natural(t)$. If we take a look at the definition (31) of the retardation matrix R_{ij} appearing in the memory kernel, we can see that it is a complicated expression traced over $\bar{\rho}(u)$. After the Markovian approximation, however, the time integral over R_{ij} is replaced by the diffusion tensor D_{ij} defined in Eq. (14), which is a complicated expression traced over $\bar{\rho}(t)$. Thus, in the Markovian approximation, we are assuming that the density, at each time t, relaxes infinitely rapidly to $\bar{\rho}(t)$. If we now take a look at the standard definition of $\bar{\rho}$, namely Eq. (8), we can see that (as discussed for the case of fluid mechanics in Grabert (1982)), this assumption corresponds to a *local equilibrium approximation*.

It is worth briefly recalling here in which way the memory term generates (in the Markovian limit) irreversible dynamics (following Zeh (2007, pp. 62-65)). The relevant information present initially is, by the operators appearing in the memory kernel, transformed into irrelevant information. This initially formed irrelevant information corresponds to so-called "doorway states" (for example two-particle correlations). The subsequent application of the orthogonal dynamics propagator $G(s,t)$ (see Appendix) transports this irrelevant information deeper into the irrelevant channel (for example by creating many-particle correlations). Due to the depth of the irrelevant channel (in a system with a large number of particles), it takes an almost infinitely long time (recurrence time) for the information to come out of the irrelevant channel again, ensuring that the "irrelevant information" is indeed irrelevant for the dynamics of the relevant degrees of freedom. The Markovian approximation in particular assumes that the relevant variables evolve so slowly that they do not change during the time it takes for the irrelevant information to move from the doorway states into the irrelevant channel, ensuring that one can effectively assume that there never is irrelevant information and that we can therefore write a closed dynamics for the relevant degrees of freedom.

Myrvold (2021, pp. 181-186) discusses this issue in a very similar way, although not explicitly based on Grabert's MZ formalism. Following Penrose (1970), he introduces as the *Markovian postulate* the assumption that "for the purposes of predicting the future macrostate, you can replace ρ with a density function $\bar{\rho}$ that yields the same probabilities for the macrovariables $\{F_1, ...F_n\}$, but which is smoothed out over surfaces that agree on the values of the macrovariables (i.e. $\bar{\rho}$ is a function of the macrovariables $\{F_1, ...F_n\}$)” (Myrvold, 2021, p. 185). Apart from the fact that his discussion is not explicitly based on the MZ formalism (and that he uses F_i rather than A_i for the relevant variables), the physical ideas expressed in this quote from Myrvold are almost identical to those used in Grabert’s approach (up to the fact that $\bar{\rho}$ is assumed to be a function of the relevant variables that satisfies a macroequivalence condition). This further supports the conclusions of Section (5) regarding how to accomodate Myrvold’s understanding of probability in the MZ formalism. Moreover, we can see how the Markovian approximation is one of the crucial steps in deriving irreversible equations of motion. This can be seen most clearly by considering examples where it is *not* satisfied. Myrvold (2021, pp. 185–186) mentions that the Markovian postulate is not exceptionless and cites as an example the spin echo experiment (Hahn, 1950), where precessing nuclear spins evolve from an apparently random to an ordered distribution. While this experiment has provoked much discussion in the philosophical literature (see, e.g., Ridderbos and Redhead (1998) and Ridderbos (2002)) it is a relatively special setup taking place in a highly controlled laboratory environment. This might appear to make such exceptions largely irrelevant for everyday physics. However, the Markovian approximation can fail to be accurate also in much more common systems, namely in glasses.

An application of the MZ formalism that is very important in physics but essentially ignored in philosophy of physics is the derivation of mode coupling theory (MCT). This method is used to model the behavior of glassy systems (Das, 2004). Roughly speaking, glasses form when particles in a dense undercooled liquid get trapped such that they cannot move to their equilibrium positions in a crystal (“caging”). This prevents the system from reaching its equilibrium state (which would correpond to a crystal), leaving it in a disordered state with strong dependence on the history of the system (“aging”) instead.

In the derivation of MCT, one projects onto density and current, which are typical slow variables for a fluid (Janssen, 2018). This allows to derive a formally exact equation of motion for the density correlator ϕ_q involving memory effects. Instead of just dropping them completely, one makes a simple ansatz for the form of the memory kernel by expressing it via the time correlation of products of density modes. It might be surprising that products of density modes are among the irrelevant variables given that we have chosen the density as a relevant one. This is a consequence of the fact that, in the Hilbert space of dynamical variables, A corresponds to a different direction than A^2 (where A is an observable) (Zwanzig, 2001, p. 151). Nevertheless, if A is slow, it is of course not unlikely that A^2 is also slow. This is precisely what happens in MCT (Kawasaki, 2009, p. 6): Since the irrelevant variables (which include quadratic density fluctuations) are also slow, they cannot be ignored, such that the final equation of motion also contains memory. (In Zeh’s terminology: the system remains in a doorway state.) For small couplings, ϕ_q goes to zero for $t \to \infty$, which means that an initial density perturbation vanishes after a sufficiently long time. However, for larger coupling constants, ϕ_q remains finite at all times, and the liquid does not go to equilibrium (Götze, 1998, p. 878). The system has undergone a transition to a nonergodic state (Fuchizaki & Kawasaki, 2002).

Hence, we have to ask ourselves why and under which conditions a Markovian approximation is possible. This, it turns out, can be a very difficult problem that continues to be of interest in physical research. Spohn (1980) has provided a detailed review of of Markovian limits (which come in different forms) and notes in particular that the so-called *hydrodynamic limit* (where one considers large length and time scales) "is poorly understood" (1980, p. 571). This limit is closely related to the crucial assumption of local thermal equilibrium (LTE), which is required, e.g., for deriving Fourier's law of heat conduction. Proving that a system stays in LTE is possible for stochastic dynamics, but much more challenging in the Hamiltonian case due to the difficulty of proving ergodicity (Bonetto et al., 2000, p. 130). (Note the crucial role of ergodicity here, which was also important in the mode-coupling theory example. In fact, nonergodicity may even prevent one-dimensional hard-rod systems from reaching equilibrium, depending on how "equilibrium" is understood (te Vrugt, forthcoming).) Bonetto et al. (2000, p. 128) therefore concluded: "There is however at present no rigorous mathematical derivation of Fourier's law [...] for any system (or model) with a deterministic, e.g. Hamiltonian, microscopic evolution." More recent discussions can be found in Michel et al. (2006) and Dhar and Spohn (2019). Even more recently, Tóth (2022) has investigated via computer simulations whether pseudo-irreversibility is present in a closed many-body system. While the answer turned out to be yes, the precise relaxation behavior was not diffusive as expected, e.g., from Fourier's law. If stochasticity (as provided by external perturbations) is really required here, one would of course have a strong basis for a new type of interventionism. However, progress has been made (Bricmont & Kupiainen, 2007; Michel et al., 2005), such that it is certainly too early to draw such a conclusion.

Independent of these developments, related issues have also been discussed in philosophy. In a recent discussion of open problems in the foundations of thermodynamics, Myrvold (2020) has argued that the real issue is not the time-reversal invariance of microscopic dynamics. Moreover, it is no mystery why, in cases where we consider a system that is a subsystem of a larger system or where we only consider a subset of a system's degrees of freedom, a system tends to forget its past in the sense that different past states are compatible with the same present state. (Such a "forgetting of the past" is what happens in equilibration, since the equilibrium state is independent of the initial state.) The real mystery, Myrvold argues, is why it is not the case that a given present states is not compatible with different future states, i.e., why we can predict the future of a system if we know only the values of a few macroscopic variables. In short, Myrvold poses as the main puzzle the explanation of autonomous macrodynamics (such that this issue also connects Myrvold's and Robertson's discussions). A good illustration would be a simple one-dimensional harmonic oscillator, fully described by its position x and momentum p. Assume that we only know the position x. It is clear that the same present value of x is compatible with different past values of x. However, it is equally clear that the present value of x is compatible with different future values of x. The interesting question is then why the situations we typically encounter in statistical mechanics are different from this harmonic oscillator. The answer has to do with the Markovian approximation. In a many-particle system, the irrelevant information typically remains in the irrelevant channels (essentially) forever, whereas in the case of the harmonic oscillator, the "irrelevant" information (the information about p) affects the relevant variable x directly and immediately. Hence, it is the Markovian approximation that we should be concerned with.

Note that the connection between irreversibility and the Markovian approximation changes if the microdynamics is not Hamiltonian. For example, if we apply the MZ formalism to a system with dissipative microscopic dynamics (an example would be the derivation of Green-Kubo relations for chiral active matter by Han et al. (2021)), then the organized drift v_i, which is always reversible in the Hamiltonian case (Grabert, 1982, p. 43), can already describe equilibration. Usually, this occurs because one applies the MZ formalism to a "microdynamics" that is already coarse-grained, such as the Langevin equations describing the motion of colloidal particles in a fluid providing friction.

However, also in these applications we generally have memory and noise terms, such that "Why is there autonomous macrodynamics?" remains an interesting question also in this context. Again, we can consider DDFT as an example, which is an autonomous equation for the one-body density of a fluid that is (usually) derived by coarse-graining the Langevin equations (Marini Bettolo Marconi & Tarazona, 1999). If we seek to explain the irreversibility of DDFT, our problem is indeed not one of time-reversibility, since the Langevin equations are already time-asymmetric. (Note, though, that using an equation of the form (13) in DDFT tends to accelerate relaxation processes (Kawasaki, 2006, p. 250), i.e., even though the system would approach equilibrium anyway because it is damped, it does so faster if we make an additional Markovian approximation). A formally exact alternative to DDFT, known as power functional theory (PFT), does, however, contain memory, and this memory is relevant for at least some physical effects (see Schmidt (2022) for a review). Moreover, a version of MCT has been derived also for overdamped colloidal particles (Szamel & Löwen, 1991). Thus, where autonomous macrodynamics comes from and whether it exists is an interesting problem also in the case of irreversible microscopic dynamics. The fact that this issue persists even in the absence of microscopic reversibility also further support's Myrvold's (2020) observation that we are facing here a problem that is distinct from the issue of time-reversibility.

8 Irreversibility (c): Arrow of time

We now turn te Vrugt's (2021) fifth problem, namely the arrow of time. The discussions in Wallace (2011) and Robertson (2020) suggest to analyze this problem based on the idea of "forward compatibility" (see Section (3.1), a condition that is satisfied if the microscopic and the $C+$ dynamics agree, and that (according to the simple dynamical conjecture) is satisfied for simple initial densities. These then explain the arrow of time. Initial state assumptions are the other ingredient (apart from the Markovian approximation) required for obtaining irreversible dynamics, and explaining them is also a considerable challenge (as emphasized in the physical literature by, e.g., Spohn (1980, p. 570) and Zwanzig (2001, p 197)). In this section, I use the MZ formalism first for a mathematical analysis of the idea of forward compatibility and then for a conceptual analysis of the question when to impose the simplicity condition on the density. For the latter question in particular, Grabert's formalism is interesting since its applications to relaxation experiments are based on the assumption that such experiments start with a simple initial state (an assumption that Wallace (2011) criticizes).

Recall that the $C+$ dynamics works by starting from an initial density $\rho(s)$, projecting it onto $\bar{\rho}(s)$, evolving it forwards in time for a small time interval Δt using the microdynamics U, applying the projection $P^\dagger$ again, evolving it forwards again and so on. Due to Eq. (11), we can then assume at each time t that the density at time $t - \Delta t$ was of the relevant form (since at this time we have applied the projection operator to eliminate every other part of the density). As shown in Section (3.2), the assumption that $\rho(s) = \bar{\rho}(s)$ allows to set $f_i(t, s) = 0$. Let us use the fact that $\rho(t - \Delta) = \bar{\rho}(t - \Delta t)$ if ρ is evolved via the $C+$ dynamics. Then, Eq. (12) gives (setting $s = t - \Delta t$)

$$\dot{a}_i(t) = v_i(t) + \int_{t-\Delta t}^{t} du R_{ij}(t, u) a_j^\natural(u). \tag{19}$$

Hence, in the $C+$ dynamics, we can calculate $\dot{a}_i(t)$ by using an extremely short memory kernel (the memory integral only covers a time Δt). There are two assumptions we have to make in order for the $C+$ result (19) to agree with the exact dynamics given by Eq. (12):

1. The memory kernel has to fall off on a very short timescale (namely Δt), such that it does not matter that we have eliminated most of the memory integral.
2. The mean random force f_i has to vanish.

If we compare these two assumptions to the two approximations we have made in Section (3.2) to arrive at the irreversible dynamic equation (13), we can see that they are exactly the same. A rapidly decaying memory kernel implies Markovian dynamics, and a vanishing mean random force is the result of $\delta\rho(s) = 0$. This result teaches us two important lessons:

1. The simplicity of the initial density ($\rho(s) = \bar{\rho}(s)$) is required for forwards-compatibility, as suggested by the simple dynamical conjecture.
2. A simple initial state is not sufficient for forward compatibility, as it does not by itself allow for a Markovian approximation. In addition to a condition on the initial state (simplicity) to solve the fifth problem, we also require a condition on the dynamics (quickly relaxing memory kernel) to solve the fourth problem.

In principle, this result should not be surprising. For a Hamiltonian system where the recurrence time is very short, there is obviously no way to get irreversible dynamics by just imposing the "right" initial condition. (Wallace (2015, p. 292), in his discussion of the MZ formalism, also notes that one requires both a time-symmetric constraint on the dynamics and a constraint on the initial state.) Since the Markovianity condition has already been discussed in Section (7), we can now turn to the other one ($\rho(s) = \bar{\rho}(s)$), which fixes the (thermodynamic) arrow of time.

Conceptually (based on the idea of "irrelevant information channels" discussed in Zeh (2007) and briefly reviewed in Section (7)), the symmetry breaking provided by the assumption $\rho(s) = \bar{\rho}(s)$ can be understood as follows: At the initial time, there is no irrelevant information ($\rho = \bar{\rho}$), and irrelevant information generated subsequently goes into the irrelevant channel and therefore does not affect the relevant dynamics. If we time-reverse this process, then the irrelevant information would come back out of the irrelevant channel and become relevant (and thus affect the time evolution of the macroscopic variables). Therefore, the irrelevant information is irrelevant only for predictions, but not for retrodictions. Moreover, the time evolute of a simple distribution is not simple since irrelevant information is created from the relevant one. Evolving the system backwards from the time s at which we imposed simplicity, the relevant information (all that there is at time s) is also transferred into the irrelevant channel. Suppose now that we had imposed the condition $\rho = \bar{\rho}$ at the end rather than at the beginning of the process that we wish to study. Then, irrelevant information has to be present prior to the end and has to transform into relevant information during the time evolution (and therefore has to affect the macroscopic dynamics). Consequently, the macroscopic dynamics is, in this case, affected by microscopic many-particle correlations. This is precisely what happens both in simulations where anti-thermodynamic behavior is observed (such as the ones by te Vrugt et al. (2021b), who artificially generated a highly correlated initial state for this purpose) and in real spin systems (Micadei et al., 2019) where anti-thermodynamic behavior arises due to initial correlations relevant for the subsequent time evolution.

Regarding the problem of symmetry breaking, Wallace (2011) notes that a simple density is compatible not only with the forward-dynamics, but also with the backward dynamics induced by a given coarse-graining procedure. The problem is that, while the forward coarse-grained dynamics is usually accurate, the backward coarse-grained dynamics is not. Moreover, the forward time evolution of a simple distribution is not simple. Hence, simplicity can only be imposed once. He then discusses two choices for the time at which it is imposed, namely

1. at the beginning of the process that one wishes to study.
2. at the beginning of time.

Wallace (2011, p. 21) relates the first option to Jaynes' objective Bayesian approach, while Robertson (2020, pp. 559 – 560), who discusses the same options, relates it to the practice of actual physics. Wallace then quickly dismisses option 1 based (among other things) on the argument that it would imply anti-thermodynamic behavior before the start of the process. Option 2, in contrast, ensures that problematic backward coarse-graining is not possible. Hence, this option should be chosen for explaining thermodynamic irreversibility. Robertson (2020, p. 560) notes that there will not be huge empirical differences between the predictions both options lead to.

While imposing simplicity at the beginning of time is indeed a reasonable way of explaining the universe's arrow of time, Wallace is in fact too quick with option 1. As discussed by Grabert (1978, p. 492), the assumption that Eq. (8) gives the initial condition for ρ is satisfied if the system starts in a state of constrained equilibrium, where (due to the application of external forces) the values of the macrovariables are forced to assume certain values. The microscopic degrees of freedom then relax towards the state that maximizes the system's entropy with respect to the macrosocopic constraint given by these macrovariables. At the start of the experiment (time

s), the external field is removed, and the system starts to evolve from the simple distribution that was forced upon it as an initial condition. In this context (which is quite typical for simulations and experiments), it is very reasonable to impose simplicity at the beginning of the process we wish to study.

What about the objection that this predicts anti-thermodynamic behavior before the beginning of this process? This objection assumes that the microscopic dynamics is the same before this time. However, before the beginning of this process (i.e., during the preparation of the experiment), the system was subject to external forces, and these external forces modify the Hamiltonian. Let us, following Grabert (1982, p. 29), assume that the system's own Hamiltonian is H and that the external forces h_i couple in such a way that they change the Hamiltonian to $H - h_i A_i$. Then, the system will relax to a generalized canonical state of the form

$$\rho = \frac{1}{Z} e^{-\beta(H - h_i A_i)} \tag{20}$$

with the rescaled inverse temperature β. The external forces are then switched off at the beginning of the process, and we observe how the system relaxes back to equilibrium. Evidently, (20) is a state of the form (8)[15], and we are thus justified in assuming $\rho(s) = \bar{\rho}(s)$ (simple initial distribution). Nevertheless, this simple initial distribution arose precisely *because* of normal thermodynamic behavior (relaxation to the state (20), which was the equilibrium state while the forces h_i were still present).

A more sophisticated objection would run as follows: The reason we expect a state of constrained equilibrium for systems prepared in this way is that the microscopic degrees of freedom will relax to a maximum-entropy state subject to these external constraints. This, however, is already an irreversible process. Consequently, we cannot use this assumption to explain the arrow of time. However, this is not what most physicists intend to do (the constrained-equilibrium-assumption is even applied to systems where an external drive is switched on at the beginning of the experiment, which implies non-thermodynamic behavior (te Vrugt & Wittkowski, 2019; Menzel et al., 2016)). While explaining the arrow of time presumably does require imposing simplicity at the beginning of time, it is perfectly reasonable to use option 1 if our goal is simply a quantitatively accurate description of a certain experiment. (Recall that, as Wallace (2015) has noted himself, explaining the arrow of time is far from the only aim of nonequilibrium statistical mechanics.)

What we do have to note, however, is that assuming that ρ is the objective density operator of the system under consideration - option (a) in the terminology introduced in Section (2.1) - the assumption $\rho(s) = \bar{\rho}(s)$ implies that the initial constrained equilibrium is not coarse-grained, but fine-grained.[16] The reason is that we make

[15] We have written Eq. 20 in a canonical form here, it can be transformed to the form (8) by including H in the set of relevant variables (Grabert, 1982).

[16] Note that Jaynes justifies the assumption $\rho = \bar{\rho}$ by arguing that we should choose the initial distribution in such a way that it maximizes the Shannon entropy subject to our knowledge, which is given by the macroscopic constraints (see Sklar (1995, pp. 255–258) and Frigg (2008)). This quite reasonable from the perspective of option (b) from Section (2.1), where $\rho(s)$ are our initial credences, but not helpful if we assume that ρ is the objective density operator. Note, however, that Jaynes' approach should not be understood as explaining or trying to explain irreversibility (Brown, 2017).

a “uniformity” assumption not (only) regarding $\bar{\rho}$, but regarding ρ itself. Such an assumption can be justified by arguing that the system, while it is being prepared, is in contact with the environment and thereby subject to external perturbations. These then destroy correlations and ensure that the microscopic state assumes the form (8). In other words, if we adapt option (a) and use the initial state assumption (20) at the beginning of an experiment, we have to be interventionists (at least for the preparation of the experiment, we may still use a coarse-graining-based notion of equilibration for the experiment itself).

If you do not like interventionism, you have two alternatives. First, you can read the assumption $\rho(s) = \bar{\rho}(s)$ in a more generous way as stating that the initial state ρ is such that it does not contain “special” correlations that would lead to anti-thermodynamic behavior. In this case, the term f_i in Eq. (12) should be and remain so small that it does not influence the macroscopic time evolution. This idea is particularly appealing if one (like Myrvold (2020, p. 139)) has reservations against singling out one particular time s as special. (Actually, even Wallace (2021, p. 15) has called the strict assumption $\delta\rho(s) = 0$ “overkill”.) Second, you can use option (b) (see Section (5), which has some advantages for this particular problem. Here, both ρ and $\bar{\rho}$ have an epistemic interpretation, such that the assumption that ρ starts in the form (20) does not commit us to believing that an actual quantum density operator reaches a fine-grained equilibrium form. Instead, it simply means that our initial credences are represented by a distribution of the form (20), which is a very reasonable credence function to have. In particular, this approach is reasonable in the context of classical simulations discussed in Section (5). If the initial conditions for such a simulation are determined via a probability distribution of the form (20), then we have a very direct explanation of why we can assume ρ and $\bar{\rho}$ to have this initial form.

9 Conclusion

I have discussed in detail the derivation of time-asymmetric transport equations from time-symmetric microscopic dynamics in modern versions of the Mori-Zwanzig projection operator formalism. This has allowed for a qualitative and quantitative examination of various claims from the philosophical literature related to the status of probability and irreversibility in statistical mechanics that are based on “simpler” mathematical formalisms. Regarding probability, it has been shown that one can understand the two distributions ρ and $\bar{\rho}$ (a) by combining Wallace’s and Jaynes’ approaches as the objective density operator and an information-theoretically constructed distribution or (b) based on Myrvold’s approach as the time evolute of the initial credences and the actual credences as a later time. The analysis of irreversibility has revealed (a) that coarse-graining in statistical mechanics can be justified both by the search for autonomous macrodynamics and by the limitations of human observers, (b) that justifying the Markovian approximation is a difficult problem that can exist even for irreversible microdynamics, and (c) how Wallace’s idea of forward compatibility can be accomodated within Grabert’s MZ formalism.

Appendix: Derivation of Eq. (12)

Here, I present in more detail the derivation of Eq. (12). From the Liouville equation (5), which holds in the Schrödinger picture, one can derive the Heisenberg picture equation of motion

$$\dot{A}_i = \mathrm{i}LA_i, \tag{21}$$

which (for a time-independent Liouvillian L) has the solution

$$\dot{A}_i(t) = e^{\mathrm{i}Lt}A_i, \tag{22}$$

where we write $A_i(0) = A_i$ (i.e., we assume Schrödinger and Heisenberg picture to coincide at time $t = 0$ (Balian & Vénéroni, 1985)). We can now insert the operator identity (Grabert, 1982, p. 16)

$$e^{\mathrm{i}Lt} = e^{\mathrm{i}Lt}P(t) + \int_s^t du e^{\mathrm{i}Lu}P(u)(\mathrm{i}L - \dot{P}(u))Q(u)G(u,t) + e^{\mathrm{i}Ls}Q(s)G(s,t) \tag{23}$$

with the orthogonal dynamics propagator

$$G(s,t) = \exp_R\left(\int_s^t du \mathrm{i}LQ(u)\right) \tag{24}$$

and the right-time-ordered exponential $\exp_R$ (see te Vrugt and Wittkowski (2019)) into Eq. (22) and average the result over $\rho(0)$[17]. As a result, we find an *exact* dynamic equation for the mean values a_i, which reads (Grabert, 1982, p. 19)

$$\dot{a}_i(t) = v_i(t) + \int_s^t du K_i(t,u) + f_i(t,s) \tag{25}$$

with the organized drift

$$v_i(t) = \mathrm{Tr}(\bar{\rho}(t)\dot{A}_i), \tag{26}$$

the memory function

$$K_i(t,u) = \mathrm{Tr}(\bar{\rho}(u)\mathrm{i}LQ(u)G(u,t)\dot{A}_i), \tag{27}$$

and the mean random force

$$f_i(t,s) = \mathrm{Tr}(\delta\rho(s)G(s,t)\dot{A}_i). \tag{28}$$

Up to now, we have not used the fact that $\bar{\rho}$ has the form (8). If we now choose it to have this form, we can use the relation (Grabert, 1978, p. 483)

$$-\mathrm{i}L\bar{\rho}(t) = a_j^\natural(t)\int_0^1 d\alpha e^{-\alpha a_k^\natural A_k}\dot{A}_j e^{\alpha a_k^\natural A_k}\bar{\rho}(t), \tag{29}$$

to get (Grabert, 1982, p. 33)

$$K_i(t,u) = R_{ij}(t,u)a_j^\natural(u) \tag{30}$$

[17] Since this derivation works in the Heisenberg picture, the density operator is constant and we can use its initial form $\rho(0)$ to get the correct average at all times.

with the retardation matrix

$$R_{ij}(t,u) = \int_0^1 d\alpha \, \mathrm{Tr}(\bar{\rho}(u) e^{\alpha a_k^\natural(u) A_k} (Q(u) G(u,t) \dot{A}_i) e^{-\alpha a_k^\natural(u) A_k} \dot{A}_j). \tag{31}$$

Inserting Eq. (30) into Eq. (25) gives Eq. (12).

Acknowledgements I am very grateful to Paul M. Näger for encouraging and supervising this project. I also wish to thank the participants of the thesis colloquium in theoretical philosophy - Ulrich Krohs, Oliver Robert Scholz, Peter Rohs, Ansgar Seide, and others - for helpful feedback on a previous version of this work. Moreover, I thank Raphael Wittkowski for introducing me to the MZ formalism, Michael Wilczek for interesting discussions about turbulence, Gyula I Tóth for introducing me to the problem of diffusive relaxation, and Alice Rolf for assistance with the layout.

Funding Open Access funding enabled and organized by Projekt DEAL. I thank the Studienstiftung des deutschen Volkes for financial support.

References

Albert, D.Z. (1994a). The foundations of quantum mechanics and the approach to thermodynamic equilibrium. *British Journal for the Philosophy of Science*, *45*(2), 669–677.

Albert, D.Z. (1994b). The foundations of quantum mechanics and the approach to thermodynamic equilibrium. *Erkenntnis*, *41*(2), 191–206.

Albert, D.Z. (2000). *Time and chance*. Harvard University Press.

Anero, J.G., Español, P., & Tarazona, P (2013). Functional thermo-dynamics: A generalization of dynamic density functional theory to non-isothermal situations. *Journal of Chemical Physics*, *139*(3), 034106.

Archer, A.J., & Rauscher, M. (2004). Dynamical density functional theory for interacting Brownian particles: Stochastic or deterministic? *Journal of Physics A: Mathematical and General*, *37*(40), 9325.

Balian, R., & Vénéroni, M. (1985). Time-dependent variational principle for the expectation value of an observable: Mean-field applications. *Annals of Physics*, *164*(2), 334–410.

Bonetto, F., Lebowitz, J.L., & Rey-Bellet, L. (2000). Fouriers law: A challenge to theorists. In A. Fokas, A. Grigoryan, T. Kibble, & B. Zegarlinski (Eds.), *Mathematical physics 2000* (pp. 128150). Imperial College Press.

Bricmont, J., & Kupiainen, A. (2007). Towards a derivation of Fourier's law for coupled anharmonic oscillators. *Communications in Mathematical Physics*, *274*(3), 555–626.

Brown, H.R. (2017). Once and for all: The curious role of probability in the Past Hypothesis. In D. Bedingham, & O. Maroney (Eds.), *Quantum foundations of statistical mechanics*. Oxford University Press. http://philsci-archive.pitt.edu/13008/

Brown, H.R., & Uffink, J. (2001). The origins of time-asymmetry in thermodynamics: The minus first law. *Studies in History and Philosophy of Modern Physics*, *32*(4), 525–538.

Chandy, A.J., & Frankel, S.H. (2010). The t-model as a large eddy simulation model for the Navier–Stokes equations. *Multiscale Modeling & Simulation*, *8*(2), 445–462.

Chen, E.K. (2021). Quantum mechanics in a time-asymmetric universe: On the nature of the initial quantum state. *British Journal for the Philosophy of Science*, *72*(4), 1155–1183.

Chorin, A.J., Hald, O.H., & Kupferman, R. (2002). Optimal prediction with memory. *Physica D: Nonlinear Phenomena*, *166*(3-4), 239–257.

Chua, E.Y.S. (2021). Does von Neumann entropy correspond to thermodynamic entropy? *Philosophy of Science*, *88*(1), 145–168.

Clarkson, C., Ellis, G., Larena, J., & Umeh, O. (2011). Does the growth of structure affect our dynamical models of the universe? The averaging, backreaction, and fitting problems in cosmology. *Reports on Progress in Physics*, *74*(11), 112901.

Das, S.P. (2004). Mode-coupling theory and the glass transition in supercooled liquids. *Reviews of Modern Physics*, *76*(3), 785–851.

Dhar, A., & Spohn, H. (2019). Fourier's law based on microscopic dynamics. *Comptes Rendus Physique*, *20*(5), 393–401.

Español, P., & Löwen, H. (2009). Derivation of dynamical density functional theory using the projection operator technique. *Journal of Chemical Physics*, *131*(24), 244101.

Farr, M. (2021). What's so special about initial conditions? Understanding the past hypothesis in directionless time. In Y. Ben-Menahem (Ed.), *Rethinking laws of nature*. Springer. http://philsci-archive.pitt.edu/19905/

Frigg, R. (2008). A field guide to recent work on the foundations of statistical mechanics. In D. Rickles (Ed.), *The Ashgate companion to contemporary philosophy of physics* (pp. 99–196). Ashgate.

Frigg, R. (2016). Chance and determinism. In A. Hájek, & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy* (pp. 460–474). Oxford University Press.

Frigg, R., & Hoefer, C. (2015). The best Humean system for statistical mechanics. *Erkenntnis*, *80*(3), 551–574.

Frisch, M. (2005a). Counterfactuals and the past hypothesis. *Philosophy of Science*, *72*(5), 739–750.

Frisch, M. (2005b). *Inconsistency, asymmetry, and non-locality: A philosophical investigation of classical electrodynamics*. Oxford University Press.

Frisch, M. (2006). A tale of two arrows. *Studies in History and Philosophy of Modern Physics*, *37*(3), 542–558.

Fuchizaki, K., & Kawasaki, K. (2002). Dynamical density functional theory for glassy behaviour. *Journal of Physics: Condensed Matter*, *14*(46), 12203–12222.

Fulde, P. (1995). *Electron correlations in molecules and solids*. Springer.

Ghirardi, G.C., Rimini, A., & Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Physical Review D*, *34*(2), 470–491.

Gibbs, J.W. (1902). *Elementary principles in statistical mechanics: Developed with especial reference to the rational foundation of thermodynamics*. Scribner's sons.

Götze, W. (1998). The essentials of the mode-coupling theory for glassy dynamics. *Condensed Matter Physics*, *1*(4), 873–904.

Grabert, H. (1978). Nonlinear transport and dynamics of fluctuations. *Journal of Statistical Physics*, *19*(5), 479–497.

Grabert, H. (1982). *Projection operator techniques in nonequilibrium statistical mechanics. Springer tracts in modern physics* (vol. 95, 1st edn.). Springer.

Hahn, E.L. (1950). Spin echoes. *Physical Review*, *80*(4), 580.

Hájek, A. (2019). Interpretations of probability. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (fall 2019 edn.). Metaphysics Research Lab: Stanford University.

Han, M., Fruchart, M., Scheibner, C., Vaikuntanathan, S., de Pablo, J.J., & Vitelli, V. (2021). Fluctuating hydrodynamics of chiral active fluids. *Nature Physics*, *17*(11), 1260–1269.

Haussmann, R. (2022). Microscopic density-functional approach to nonlinear elasticity theory. *Journal of Statistical Mechanics: Theory and Experiment, 2022*, 053210.

Hemmo, M., & Shenker, O. (2006). Von Neumann's entropy does not correspond to thermodynamic entropy. *Philosophy of Science*, *73*(2), 153–174.

Henderson, L. (2003). The von Neumann entropy: A reply to Shenker. *British Journal for the Philosophy of Science*, *54*(2), 291–296.

Huang, X., Kodama, T., Koide, T., & Rischke, D. H. (2011). Bulk viscosity and relaxation time of causal dissipative relativistic fluid dynamics. *Physical Review C*, *83*(2), 024906.

Janssen, L. (2018). Mode-coupling theory of the glass transition: A primer. *Frontiers in Physics*, *6*, 97.

Jaynes, E.T. (1975a). Information theory and statistical mechanics. *Physical Review*, *106*, 620–630.

Jaynes, E.T. (1975b). Information theory and statistical mechanics. ii. *Physical Review*, *108*, 171–190.

Kawasaki, K. (2006). Interpolation of stochastic and deterministic reduced dynamics. *Physica A: Statistical Mechanics and its Applications*, *362*(2), 249–260.

Kawasaki, K. (2009). A mini-review of structural glasses—a personal view—. *Forma*, *24*(1), 3–9.

Kawasaki, K., & Gunton, J.D. (1973). Theory of nonlinear transport processes: Nonlinear shear viscosity and normal stress effects. *Physical Review A*, *8*, 2048–2064.

Klippenstein, V., Tripathy, M., Jung, G., Schmid, F., & van der Vegt, N. F. A. (2021). Introducing memory in coarse-grained molecular simulations. *Journal of Physical Chemistry B*, *125*(19), 4931–4954.

La Caze, A. (2016). Frequentism. In A. Hájek & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy* (pp. 341–359). Oxford University Press.

Löwen, H. (1994). Melting, freezing and colloidal suspensions. *Physics Reports*, *237*(5), 249–324.

Luczak, J. (2016). On how to approach the approach to equilibrium. *Philosophy of Science*, *83*(3), 393–411.

Luczak, J. (2018). How many aims are we aiming at? *Analysis*, *78*(2), 244–254.

Maeyama, S., & Watanabe, T.H. (2020). Extracting and modeling the effects of small-scale fluctuations on large-scale fluctuations by Mori–Zwanzig projection operator method. *Journal of the Physical Society of Japan*, *89*(2), 024401.

Marini Bettolo Marconi, U., & Tarazona, P. (1999). Dynamic density functional theory of fluids. *Journal of Chemical Physics*, *110*(16), 8032–8044.

Menzel, A.M., Saha, A., Hoell, C., & Löwen, H. (2016). Dynamical density functional theory for microswimmers. *Journal of Chemical Physics*, *144*(2), 024115.

Meyer, H., Voigtmann, T., & Schilling, T. (2017). On the non-stationary generalized Langevin equation. *Journal of Chemical Physics*, *147*(21), 214110.

Meyer, H., Voigtmann, T., & Schilling, T. (2019). On the dynamics of reaction coordinates in classical, time-dependent, many-body processes. *Journal of Chemical Physics*, *150*(17), 174118.

Micadei, K., Peterson, J.P.S., Souza, A.M., Sarthour, R.S., Oliveira, I.S., Landi, G.T., Batalhão, T.B., Serra, R.M., & Lutz, E. (2019). Reversing the direction of heat flow using quantum correlations. *Nature Communications*, *10*(1), 2456.

Michel, M., Mahler, G., & Gemmer, J. (2005). Fourier's law from Schrödinger dynamics. *Physical Review Letters*, *95*(18), 180602.

Michel, M., Gemmer, J., & Mahler, G. (2006). Microscopic quantum mechanical foundation of Fourier's law. *International Journal of Modern Physics B*, *20*(29), 4855–4883.

Mori, H. (1965). Transport, collective motion, and Brownian motion. *Progress of Theoretical Physics*, *33*(3), 423–455.

Myrvold, W. (2016). Probabilities in statistical mechanics. In A. Hajek, & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy* (pp. 573–600). Oxford University Press.

Myrvold, W.C. (2011). Probabilities in statistical mechanics: Objective, subjective, or a bit of both? http://philsci-archive-dev.library.pitt.edu/8642/

Myrvold, W.C. (2012). Deterministic laws and epistemic chances. In Y. Ben-Menahem, & M. Hemmo (Eds.), *Probability in physics* (pp. 73–85). Springer.

Myrvold, W.C. (2020). Explaining thermodynamics: What remains to be done? In V. Allori (Ed.), *Statistical mechanics and scientific explanation* (pp. 113–143). World Scientific.

Myrvold, W.C. (2021). *Beyond chance and credence: A theory of hybrid probabilities*. Oxford University Press.

Nakajima, S. (1958). On quantum theory of transport phenomena: Steady diffusion. *Progress of Theoretical Physics*, *20*(6), 948–959.

Orlandini, S., Meloni, S., & Ciccotti, G. (2011). Hydrodynamics from Statistical mechanics: Combined dynamical-NEMD and conditional sampling to relax an interface between two immiscible liquids. *Physical Chemistry Chemical Physics*, *13*(29), 13177–13181.

Parish, E.J., & Duraisamy, K. (2017). Non-Markovian closure models for large eddy simulations using the mori-Zwanzig formalism. *Physical Review Fluids*, *2*(1), 014604.

Penrose, O. (1970). *Foundations of statistical mechanics: A deductive treatment*. Pergamon Press.

Ras, T., Szafarczyk, M., & Fuchs, M. (2020). Elasticity of disordered binary crystals. *Colloid and Polymer Science*, *298*, 803–818.

Rau, J., & Müller, B. (1996). From reversible quantum microdynamics to irreversible quantum transport. *Physics Reports*, *272*(1), 1–59.

Redhead, M.L.G. (1995). *From physics to metaphysics*. Cambridge University Press.

Ridderbos, K. (2002). The coarse-graining approach to statistical mechanics: How blissful is our ignorance? *Studies in History and Philosophy of Modern Physics*, *33*(1), 65–77.

Ridderbos, T.M., & Redhead, M.L.G. (1998). The spin-echo experiments and the second law of thermodynamics. *Foundations of Physics*, *28*(8), 1237–1270.

Robertson, B. (1966). Equations of motion in nonequilibrium statistical mechanics. *Physical Review*, *144*, 151–161.

Robertson, K. (2020). Asymmetry, abstraction, and autonomy: Justifying coarse-graining in statistical mechanics. *British Journal for the Philosophy of Science*, *71*(2), 547–579.

Robertson, K. (forthcoming). In search of the holy grail: How to reduce the second law of thermodynamics. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/714795

Sagaut, P. (2006). *Large eddy simulation for incompressible flows* (3rd edn.). Springer.

Schilling, T. (2022). Coarse-grained modelling out of equilibrium. *Physics Reports*, *972*, 1–45.

Schmidt, M. (2022). Power functional theory for many-body dynamics. *Reviews of Modern Physics*, *94*(1), 015007.

Shenker, O.R. (1999). Is-kTr (ρlnρ) the entropy in quantum mechanics? *British Journal for the Philosophy of Science*, *50*(1), 33–48.

Sklar, L. (1995). *Physics and chance: Philosophical issues in the foundations of statistical mechanics*. Cambridge University Press.

Spohn, H. (1980). Kinetic equations from Hamiltonian dynamics: Markovian limits. *Reviews of Modern Physics*, *52*(3), 569–615.

Szamel, G., & Löwen, H. (1991). Mode-coupling theory of the glass transition in colloidal systems. *Physical Review A*, *44*(12), 8215–8219.

te Vrugt, M. (2021). The five problems of irreversibility. *Studies in History and Philosophy of Science*, *87*, 136–146.

te Vrugt, M. (forthcoming). How to distinguish between indistinguishable particles. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/718495, arXiv:2112.00178

te Vrugt, M., & Wittkowski, R. (2019). Mori-Zwanzig projection operator formalism for far-from-equilibrium systems with time-dependent Hamiltonians. *Physical Review E*, *99*, 062118.

te Vrugt, M., & Wittkowski, R. (2020a). Projection operators in statistical mechanics: A pedagogical approach. *European Journal of Physics*, *41*(4), 045101.

te Vrugt, M., & Wittkowski, R. (2020b). Relations between angular and Cartesian orientational expansions. *AIP Advances*, *10*(3), 035106.

te Vrugt, M., Löwen, H., & Wittkowski, R. (2020). Classical dynamical density functional theory: From fundamentals to applications. *Advances in Physics*, *69*(2), 121–247.

te Vrugt, M., Hossenfelder, S., & Wittkowski, R. (2021a). Mori-Zwanzig formalism for general relativity: A new approach to the averaging problem. *Physical Review Letters*, *127*, 231101.

te Vrugt, M., Tóth, G.I., & Wittkowski, R. (2021b). Master equations for Wigner functions with spontaneous collapse and their relation to thermodynamic irreversibility. *Journal of Computational Electronics*, *20*, 2209–2231.

Tóth, G. I. (2022). Emergent pseudo time-irreversibility in the classical many-body system of pair interacting particles. *Physica D: Nonlinear Phenomena, 437*, 133336.

Von Kutschera, F. (1969). Zur Problematik der naturwissenschaftlichen Verwendung des subjektiven Wahrscheinlichkeitsbegriffs. *Synthese*, *20*, 84–103.

Wallace, D. (2011). The Logic of the Past Hypothesis. In B. Loewer, E. Winsberg, & B. Weslake (Eds.), *Time's arrows and the probability structure of the world, Harvard*. http://philsci-archive.pitt.edu/8894/

Wallace, D. (2012). *The emergent multiverse: Quantum theory according to the Everett interpretation*. Oxford University Press.

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

Wallace, D. (2021). Probability and irreversibility in modern statistical mechanics: Classical and quantum. In D. Bedingham, O. Maroney, & C. Timpson (Eds.), *Quantum foundations of statistical mechanics*. Oxford University Press. arXiv:2104.11223

Walz, C., & Fuchs, M. (2010). Displacement field and elastic constants in nonideal crystals. *Physical Review B*, *81*(13), 134110.

Wittkowski, R., Löwen, H., & Brand, H. R. (2012). Extended dynamical density functional theory for colloidal mixtures with temperature gradients. *Journal of Chemical Physics*, *137*(22), 224904.

Wittkowski, R., Löwen, H., & Brand, H. R. (2013). Microscopic approach to entropy production. *Journal of Physics A: Mathematical and Theoretical*, *46*(35), 355003.

Zeh, H. D. (2007). *The physical basis of the direction of time* (5th edn.). Springer Verlag.

Zwanzig, R. (1960). Ensemble method in the theory of irreversibility. *Journal of Chemical Physics*, *33*(5), 1338–1341.

Zwanzig, R. (1973). Nonlinear generalized Langevin equations. *Journal of Statistical Physics*, *9*(3), 215–220.

Zwanzig, R. (2001). *Nonequilibrium statistical mechanics*. Oxford University Press.