Independent Evidence in Multi-messenger Astrophysics

Jamee Elder

February 12, 2024

Abstract

In this paper I discuss the first "multi-messenger" observations of a binary neutron star merger and kilonova. These observations, touted as "revolutionary," included both gravitational-wave and electromagnetic observations *of a single source*. I draw on analogies between astrophysics and historical sciences (e.g., paleontology) to explain the significance of this for (gravitational-wave) astrophysics. In particular, I argue that having independent lines of evidence about a target system enables the use of argumentative strategies—the "Sherlock Holmes" method and consilience—that help overcome the key challenges astrophysics faces as an observational and historical science.

1 Introduction

In August 2017, the LIGO-Virgo Collaboration detected gravitational waves from a binary neutron star inspiral.[1] This highly-anticipated event, dubbed "GW170817," is notable for its electromagnetic counterparts, including a short gamma-ray burst (sGRB), "GRB 170817A", and an optical transient, "AT 2017gfo."[2] This was the first joint detection of gravitational and electromagnetic radiation from a single source, as well as the first (and so far *only*) multi-messenger observation to include gravitational waves.

Multi-messenger astrophysics is done by combining information from two or more of the four cosmic "messengers": electromagnetic radiation, gravitational radiation, neutrinos, and cosmic rays (Bartos and Kowalski 2017, 1).[3] These messengers provide independent lines of evidence about the event that produced them, in ways to be elaborated in this paper.[4] In the case of GW170817, multiple observations of its source allowed for a swath of new insights about astrophysical mechanisms (e.g., r-process nucleosynthesis), fundamental physics (e.g., new constraints on deviations from general relativity), and cosmology (e.g., new, independent measurements of the Hubble constant).[5] In this paper I will analyze the epistemic situation and epistemic resources of gravitational-wave astrophysics in light of these new, multi-messenger observations.

The paper will proceed as follows. In section 2 I briefly summarize the multi-messenger observations of the astrophysical event corresponding to GW170817. In section 3 I discuss the epistemic challenges of (gravitational-wave) astrophysics, drawing on analogies with other historical sciences. I then describe two argumentative strategies that are important in this context: the unification of "traces" by a common cause—the

1. A compact binary merger has three main stages: the inspiral, where the components circle one another in close, decaying orbits; the merger, where they plunge together and coalesce; and the ringdown, where the remnant settles into a stable state. Gravitational waves are emitted at each stage.

2. For details of the electromagnetic observations, see the summary provided in sections 2.2-3.4 of Abbott et al. (2017e). See especially tables 1-6 for a summary of observations, including references for these observations. See also Abelson (2022) for a detailed summary of these observations and their specific connections to aspects of the kilonova model, highlighting some of the specific contributions by various groups included in the above-mentioned tables).

3. See also Mészáros et al. (2019) for a review of recent multi-messenger observations and their significance.

4. In particular, I will argue that they are independent to the extent that the use of them as evidence about the source system relies on different auxiliary assumptions (due to, e.g., the different emission mechanisms of different messengers, and different kinds of measuring instruments used to detect them).

5. Note that "r-process" refers to the "rapid neutron capture process", a set of nuclear reactions responsible for the creation of many heavy elements.

"Sherlock Holmes" method—and consilience. In section 4 I then provide concrete examples of both the Sherlock Holmes and consilience strategies based on observations of the source of GW170817.

This paper carefully teases apart the different kinds of arguments made possible in the case study, due to the availability of multiple, independent lines of evidence. In doing so, it demonstrates that the revolutionary nature of multi-messenger astrophysics can be understood in terms of the argumentative strategies that can be deployed with multi-messenger evidence.

These argumentative strategies crucially depend on having a diverse set of traces, providing independent lines of evidence about the target system. However, like gravitational-wave astrophysics, multi-messenger astrophysics is in its early development, so the methodological "revolution" must be regarded as a promissory note, albeit one that comes with promising initial results.

In making this case for multi-messenger astrophysics, I am also contributing to the overall case in favor of the reliability of astrophysics, begun by e.g., Anderl (2016), Jacquart (2020), Boyd (2018b), and Wilson (2017).[6] This work stands in opposition to pessimistic accounts (e.g., Hacking (1989)) according to which astrophysics is methodologically and epistemically impoverished relative to traditional experimental sciences. In recent years, a number of philosophers have taken up work in this vein (see e.g., Boyd et al. (2023)), including work by Patton (2020), Abelson (2022), and Elder (2023) on gravitational-wave astrophysics specifically. Overall, this paper makes the case that multi-messenger astrophysics expands the epistemic resources of astrophysics, and in doing so helps overcome some of the key challenges it faces as an observational and historical science.

2 Observations of a Binary Neutron Star Merger and Kilonova

The observation of the source of GW170817 was a global effort, involving thousands of astrophysicists observing across multiple "messengers" (gravitational and electromagnetic radiation) and many different "windows" within the electromagnetic

6. In particular, my discussion of the Sherlock Holmes strategy builds off of Anderl (2016)'s emphasis on the prevalence of this method in astronomy and astrophysics.

spectrum. The electromagnetic emission can be broadly classified into three components: the sGRB, the kilonova (i.e., an ultraviolet, optical, and infrared transient), and the delayed X-ray and radio counterpart.[7] Together, they offer a "comprehensive, sequential description of the physical processes related to the merger of a binary neutron star" (Abbott et al. 2017e, 27).

On 17 August 2017, a sGRB ("GRB 170817A") was detected by the *Fermi* Gamma-ray Burst Monitor (*Fermi*-GBM).[8] Analysis of the LIGO-Hanford data then identified a gravitational wave candidate ("GW170817") with a merger time less than two seconds before GRB 170817A, consistent with the expectation of gamma-ray emission within seconds after the merger. A GCN Notice was then issued, informing the astrophysics community that a signal consistent with a binary neutron star merger was associated with GRB 170817A.[9] This announcement launched a major observation campaign.

Subsequent analysis of data from LIGO-Hanford, LIGO-Livingston, and Virgo confirmed a highly significant coincident signal across the detectors and placed strong constraints on the location of the source. See figure 1 for a visual representation of this localization. Guided by these constraints, a bright optical transient, "AT 2017gfo" (initially "SSS17a") was first discovered by the One-Meter, Two-Hemisphere (1M2H) team, within eleven hours of the gamma-ray and gravitational-wave detections (Coulter et al. 2017).

The transient was observed across a range of electromagnetic frequencies as its properties changed over subsequent weeks. Through these observations, a relatively detailed picture of the source could be pieced together. Some features, such as the rapid luminosity decline and the fact that this fading was much slower in the infrared, marked this as an unprecedented observational event, consistent with models of kilonovae (Abbott et al. 2017e, 7). Signatures of heavy element nucleosynthesis via rapid neutron-capture (r-process) were also observed, providing insights into an important astrophysical mechanism for heavy element production (Pian et al. 2017; Abbott et al. 2017b).

7. The kilonova results from isotropic ejection of r-process nuclei from the merger. These nuclei undergo radioactive decay, resulting in electromagnetic emission.

8. A detailed description of these events is given in Abbott et al. (2017e).

9. The General Coordinates Network (GCN), run by NASA, shares alerts and rapid communications about high-energy, multimessenger, and transient phenomena across the astrophysics community. For further details, see https://test.gcn.nasa.gov/docs/#what-is-gcn.

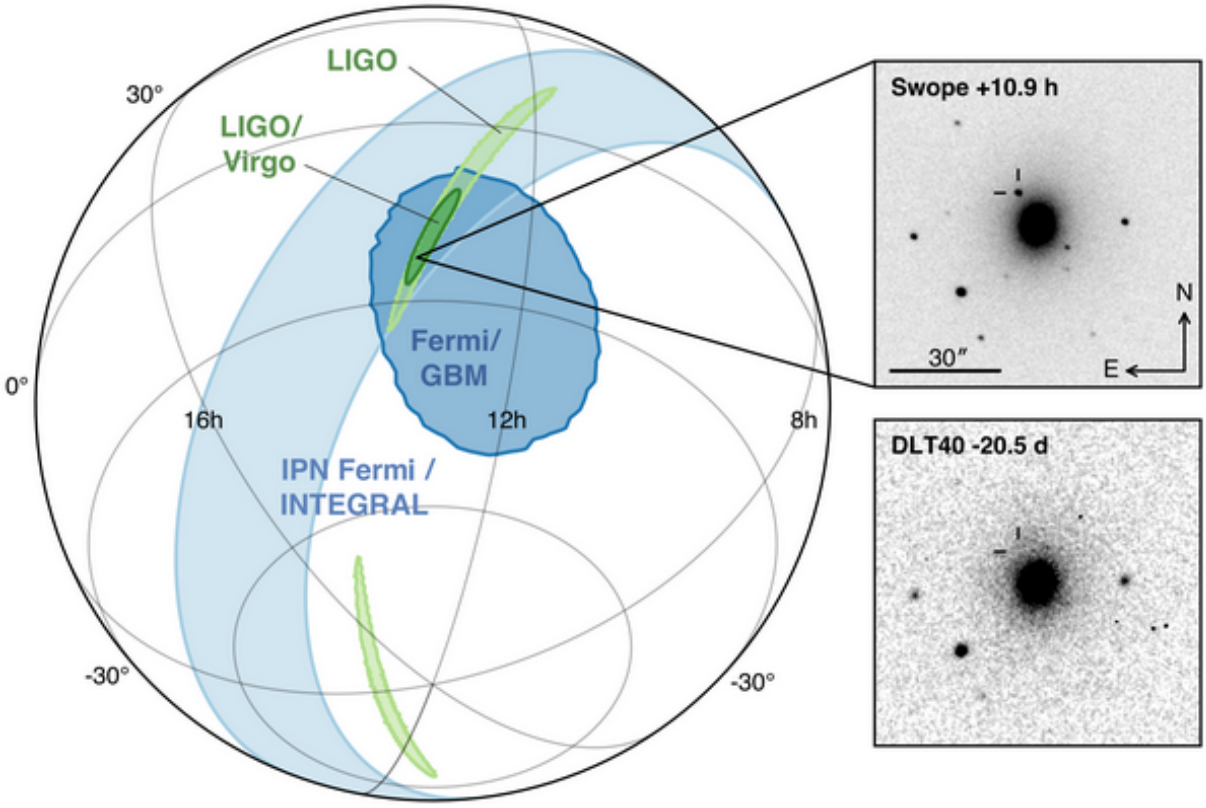

Figure 1: Image reproduced from Abbott et al. (2017e, 2).

The gravitational waves, gamma rays, and optical transient were produced by distinct emission mechanisms and originate from different periods during the evolution of the system. As such, they provide independent windows into different features of the source system over the period of observation.[10]

3 The Epistemic Situation of Astrophysics

The announcement of the first multi-messenger observation of a binary neutron star merger was met with great excitement and described as "revolutionary."[11] But what makes this event revolutionary, compared to earlier gravitational-wave detections?

In what follows, I distinguish between the *epistemic situation*, characterized by the challenges that scientists face in producing epistemic goods in a particular context, and *epistemic resources*, including the knowledge, capacities, sources of evidence, and technologies available to scientists in a given context (Currie 2018, 15). The above question about the significance of multi-messenger observations can be couched in this language as follows: how do the epistemic resources of multi-messenger astrophysics differ from those of single-messenger astrophysics? And how is this important in meeting the challenges imposed by the epistemic situation of (gravitational-wave) astrophysics?

In broad strokes, the answer to both questions lies in having multiple, independent lines of evidence about the same target system. This includes providing *complementary* evidence, which can be combined to gain new insights about the target; and *convergent* evidence—independent lines of evidence that support the same conclusion. In section 3.1 I argue that observational data in astrophysics plays the role of "traces" in historical sciences. I then characterize the epistemic situation of gravitational-wave astrophysics in terms of the challenge of gathering diverse traces that are informative about the target. In section 3.2, I describe two key argumentative strategies that have been emphasized by philosophers of historical sciences as characteristic of the reasoning in these fields. My discussion of these methods is motivated by their usefulness in illustrating the structure of argumentation in my case study, and illuminating similarities (and differences) in the epistemic situations of astrophysics and other historical sciences.

10. The significance of the different times of emission is discussed by Abelson (2022).
11. See, e.g., Holz (2017).

3.1 Traces and Observations

Astrophysics is an observational science; the empirical evidence it relies on are observations of naturally occurring phenomena, rather than results of controlled experiments. Astrophysics is also an historical science in that the events being observed (the target systems) are usually located in the distant past. For example, at a distance of approximately $40Mpc$ or $130Mly$, the events that produced GW170817 occurred approximately 130 million years ago. Astrophysical observation thus resembles the analysis of geological strata; just as deeper layers in strata provide insights into the increasingly distant past, observations of increasingly distant systems are observations of the more and more remote past.

The observational and historical nature of astrophysics leads to strong analogies with sciences like archaeology, geology, and paleontology (Anderl 2016). Like other historical scientists, astrophysicists cannot intervene on or manipulate the target systems that they study. Instead of manufacturing empirical evidence in a carefully controlled experimental system, they must rely on the naturally occurring evidence provided by the universe. This feature of historical sciences is often thought to leave them methodologically and epistemically impoverished compared to their experimental counterparts (Cleland 2002; Hacking 1989, 475). However, differences in epistemic situation do not automatically equate to epistemic inferiority; what matters is whether scientists have adequate epistemic resources to meet the specific challenges of their epistemic situation.

Recent philosophical analyses of the methods of historical sciences focuses on "traces", where a trace x is a downstream consequence of some past event or entity y, such that we can learn about y by studying x (see e.g., Cleland (2002), Currie (2018, ch.3)). Thus traces provide evidence about past target systems. As Currie (2018, 70) emphasizes, traces are evidence in virtue of background knowledge connecting the trace to the past event—we need "midrange theories" according to which the trace is evidentially relevant to the target. Midrange theories identify "dependency relations" between the target and the trace. Dependency relations are ways that the state of the trace depends on the state of the target, or vice versa. For Currie, this is to be understood as a probabilistic notion: a variable v_1 is "minimally dependent" on another variable v_2 when the fact that v_2 takes some value (or range of values) affects the probability that a variable v_1 takes on some value (or range of values) (74). Exploiting dependencies between past and present, using a midrange theory, allows for inferences to

be made about the target on the basis of traces.

The paradigm cases of traces in the historical sciences are physical entities such as fossils, artifacts, strata. In astrophysics, "messengers" are natural candidates for traces. However, messengers themselves make for awkward evidence; unlike, say, fossils, they cannot be collected and saved for further inspection. In astrophysics, the evidential currency is not the messengers themselves, but rather the records of interactions between detectors and these physical traces—in short, data. Data stand in dependency relations with the target. They are thus able to play the right evidential role to be used as traces. However, nothing in this paper rests on whether data are themselves traces. What is important is that these data count as evidence, relative to midrange theories describing dependencies between the data and the target.

The picture just described has some similarities to Dudley Shapere (1982)'s account of astrophysical observation. On this view, a (direct) observation is one where, according to our best science and background knowledge, information is transmitted, without interference, from the source to an appropriate receptor (492). Shapere breaks the background knowledge down into the "theory of the source," "theory of the transmission," and the "theory of the receptor". These are analogous to Currie's "midrange theories," since both concern the connection between the target system and the trace.

This picture also has clear connections to recent work in the philosophy of data, such as Leonelli's "relational" view of data, and Bokulich and Parker's "pragmatic representational" (PR) view of data. According to Leonelli (2016), what counts as data is relative to particular research situations. Data are 'any product of research activities [...] that [are] collected, stored, and disseminated *in order to be used as evidence for knowledge claims*' (77, emphasis in original). Bokulich and Parker (2021, 31) similarly emphasize that 'data do not have fixed evidential value', but are rather deemed fit for a given purpose relative to the 'interests, background knowledge, and other resources available to researcher.' Using data as *evidence* on these views depends on the user's ability to exploit relationships between the data and the target.

Applied to astrophysics, the core insight from these accounts of traces and data is that using observational data as evidence about past events (e.g. a binary neutron star merger) is theory- and context-dependent. The data is evidence about a target system relative to the epistemic resources that a scientist can bring to bear in a particular epistemic situation. In astrophysics, this includes: models of the source system and

emission mechanisms; theoretical descriptions of the radiative processes governing the transmission from source to measuring instruments; and models of the measuring process.[12] The independence of different messengers in multi-messenger astrophysics comes from the lack of overlap in these various models of the source, transmission, and detection processes.

In gravitational-wave astrophysics the most important "midrange theories" are applications of general relativity. This theory governs the dynamics of the source and its gravitational wave emission, the propagation of these waves through spacetime, and the eventual interaction of these with the "test masses" of the receiver (mirrors in the gravitational wave interferometers). Interpreting the LIGO-Virgo data relies on having a template library of accurate gravitational waveforms worked out in advance, which in turn relies on modeling the dynamics of a binary systems in the "dynamical strong field regime" (i.e., where both high velocities and strong gravitational fields come into play). There are no exact analytic solutions to this problem, so instead a range of modeling approaches are used—both to model the binary system, and to extract information about the gravitational waves emitted. In particular, numerical relativity simulations are thought to provide the most accurate insights into the late-time behaviour of these binaries. The first successful simulations of binary black hole mergers were performed in 2005 (Pretorius 2005; Campanelli et al. 2006; Baker et al. 2006). These models capture dependency relations between the gravitational wave data and the target system. For example, the frequency f and its time derivative $\dot{f}$, as measured by the gravitational-wave interferometer, are related to the masses of the binary components (see equation (1), below). Additionally, the amplitude of the gravitational waves encodes information about the distance to the binary (see section 4.2.1). Theoretical breakthroughs in modeling compact binary systems have made it possible to exploit these and other dependency relations between the LIGO-Virgo data and the source system. This illustrates how the expansion of epistemic resources, in the form of midrange theories, expands what traces can be used as evidence.

Gravitational-wave astrophysics faces slightly different challenges than standard *electromagnetic* astrophysics does. To begin with, gravitational waves are much (much!)

12. A contemporary view of measurement as a "model-based inference" has recently emerged in the work of van Fraassen (2008), Tal (2012, 2013, 2016, 2017), and Parker (2017). This provides an insightful lens for understanding the measuring processes of LIGO-Virgo. Some relevant philosophical analyses of the role of models in the LIGO-Virgo data-analysis are Elder (2023, 2020) and Patton (2020).

harder for us to detect. Devices that detect and utilize electromagnetic radiation abound (e.g., radios, x-rays, microwaves). In contrast, the detection of gravitational waves is so challenging that there was significant doubt it would ever be accomplished (Kennefick 2007).

Electromagnetic astrophysics is better established than gravitational-wave astrophysics. First, a much greater number and variety of observations are currently possible; astronomers can observe a full spectrum of frequencies with a range of different instruments. In contrast, gravitational-wave astrophysics is currently limited to three detectors capable of detecting plausible signals. These are all extremely similar and sensitive to the same frequency band. The close proximity and orientation of the two LIGO detectors adds further redundancy to the network, since this reduces information about the polarization of the gravitational waves. While this redundancy is in some ways unfortunate, it was a deliberate choice to prioritize confidence in detections. Here there is an important epistemic tradeoff between collecting more information about the signal (i.e., diversifying traces) and increasing the security of the detection claim. The two LIGO detectors should record very similar responses to passing gravitational waves and this provides an important consistency check on the individual results of each detector.

Our level of theoretical understanding of Maxwell's theory of electromagnetism also far exceeds our comprehension of the theoretical apparatus of general relativity. Likewise, our application of electromagnetism to a variety of material systems constitutes a mastery over electromagnetism that we lack in the case of gravitational phenomena. Through these applications, Maxwell's theory enjoys a far greater level of empirical entrenchment than Einstein's.

However, one benefit of gravitational radiation over electromagnetic radiation is the level of interaction and interference between the incident signal and intervening matter. Gravitational waves are only weakly absorbed and scattered by interactions with such matter. Electromagnetic waves are absorbed and scattered to a much greater extent, so astronomers must correct for the effect of intervening matter (including Earth's atmosphere). That said, the stronger interactions of electromagnetism with intervening matter can also be a benefit. For example, absorption lines and gravitational lenses are valuable sources of information about the intervening entities.

3.2 Sherlock Holmes, Smoking Guns, and Consilience

Astrophysics, like other historical sciences, relies on traces, which provide records of the state of distant target systems at past times. In astrophysics, there are four main classes of trace: photons, neutrinos, cosmic rays, and gravitational waves. Multi-messenger events are rare, but in such special cases the different messengers provide independent lines of evidence about the system that produced them.

A trace is "evidence" concerning a hypothesis when the trace bears on whether that hypothesis is true. However, we have seen that traces are only evidence relative to midrange theories. For a trace to be evidence, it must stand in a dependency relationship with the target system, making the state of the trace informative about some feature of the target. Midrange theories describe the "auxiliary conditions" under which these dependency relations hold.

"*Independent* evidence" refers to cases where we have separate pieces of evidence, E_1 and E_2, such that their status as evidence depends on different auxiliary conditions holding. I will take a "line of evidence" to refer to the evidence itself plus the auxiliary conditions necessary for it being evidence.[13] *Independent* lines of evidence are then lines of evidence where there is no overlap in auxiliary conditions. Modeling a line of evidence in astrophysics will thus involve modeling the detector, the propagation of the messenger from the source, and the source itself. Independence (or partial independence) between these lines of evidence can then be bought in various ways: different measuring processes (e.g., gravitational wave interferometers vs. telescopes), different messengers (e.g., gravitational waves vs. photons), and different emission mechanisms (e.g., motion of massive bodies vs. radioactive decay of heavy nuclei).

Strictly speaking, two lines of evidence might never be fully independent because there are always sufficiently basic physical conditions or theoretical facts whose holding is a necessary condition for both E_1 and E_2 to count as evidence. In practice, we are less concerned about overlap of assumptions that are sufficiently empirically entrenched. We *are* concerned about overlap in assumptions about auxiliary conditions that present plausible sources of error for both pieces of evidence. In such cases, both E_1 and E_2 might be evidence for a given hypothesis on the assumption that this shared auxiliary condition holds, but not otherwise. Both of the lines of evidence could then fail to be reliable for the same reason.

13. See Boyd (2018a) for a related account of "enriched" evidence.

In the remainder of this section, I discuss two key argumentative strategies used in the historical sciences, which rely on combining independent lines of evidence—the "Sherlock Holmes" method, and consilience. This exposition supports the analysis of my case study in section 4, and ultimately helps illuminate the benefits of multi-messenger astrophysics.

Trace-based accounts of methodology in historical sciences often emphasize the unification of diverse traces by the positing of a common cause (e.g., Cleland (2002), Tucker (2011), Kleinhans, Buskes, and de Regt (2005)). A classic example is the explanation of the Cretaceous–Paleogene (K–Pg) extinction event, which included the extinction of the non-avian dinosaurs[14]. Some of the traces included the ammonite fossil record, elevated iridium levels at the K-Pg boundary, deposits of shocked quartz, and the discovery of a large crater, the Chicxulub Crater. All of these traces can be explained by a single common cause: a large asteroid impact approximately 65 million years ago (see e.g., Cleland (2002, 483)). However, without a common cause to unite them, these are unconnected features of the world.

As Cleland (2002) notes, there were many candidate explanations for the K-Pg extinction event prior to 1980. However, the discovery of elevated iridium concentrations eliminated all but two hypotheses: first, a period of increased volcanic activity; and second, an asteroid impact. These were considered the only plausible mechanisms for generating the observed iridium concentrations (high concentrations of iridium being present in both asteroids and the Earth's mantle). The discovery of the shocked quartz decided the matter between these two theories because shocked quartz is generally only found in two places on Earth: sites of asteroid impacts and sites of nuclear explosions. Thus the fossil record, the iridium concentrations, and the quartz were considered sufficient to justify acceptance of the asteroid impact hypothesis. The shocked quartz acted as a "smoking gun", deciding between the two previously viable contenders. Cleland (2002) argues that the search for smoking guns is a vital element of historical methodology. This search is the historical equivalent of devising new experiments to help distinguish between rival hypotheses in the experimental sciences. However, Cleland does not seem to take smoking guns to be the analog of "crucial" or "critical" experiments since the smoking gun is the "capstone," that "can only be judged as a smoking gun when combined with the rest of the evidence available" (Cleland 2013, 4).

14. I follow Cleland in using this example to illustrate the method of unification by a common cause, but I am not committed to this being the only or best way of analyzing the evidence in this case.

In other words, smoking guns do not generally constitute decisive evidence for one hypothesis or another *by themselves*. Below, in section 4.1, I identify the optical transient as the " smoking gun" in the case study. However, I think we should be cautious about placing too much emphasis on smoking guns. The pursuit of smoking guns may be a fruitful approach to gathering new evidence, but, as Currie puts it, the particular evidence that is designated the "smoking gun" may "play a relatively minor evidential role overall" (Currie 2018, 235-6).

The unification of traces by a common cause is sometimes referred to as a "Sherlock Holmes" methodology because historical scientists are like criminal detectives, assembling clues to crack the case (Cleland 2002; Anderl 2016). Anderl (2016) emphasizes the role of this "Sherlock Holmes" strategy in constructing models of particular events or processes in astrophysics. The idea behind this strategy is that no single piece of evidence ("clue") is sufficient to mount a convincing argument. However, in conjunction with other traces, it is possible to argue for the conclusion by pointing to a common cause that explains them all.[15] I discuss an example of the Sherlock Holmes strategy based on GW170817/GRB 170817A/AT 2017gfo in section 4.1.

Other accounts of historical methodology emphasize *consilience*, which involves exploiting multiple independent lines of evidence to obtain convergent results (e.g., Forber and Griffith (2011), Wylie (2011)). This convergence is thought to increase the "robustness" of the shared conclusion and improve trust in the convergent results. Here, robustness is a property of the results[16]

Given the variety of notions of "robustness" in the philosophical literature, it is worth briefly stating how I am using the term. First, I take robustness to be a *property* of a hypothesis or claim (as opposed to robustness *analysis* (or consilience) which is a *strategy* for providing confirmation by combining different lines of evidence)[17] A

15. Like much of scientific reasoning, the Sherlock Holmes method is a form of inference to the best explanation (IBE) or abductive reasoning—one where diverse and potentially unconnected traces are unified by inferring that they are best explained by a common cause.

16. Philosophical interest in "robustness" and "consilience", as well as related notions such as "stability", and "reproducibility"), has a rich history which I cannot do justice to here. Wimsatt (1981) provides a detailed philosophical discussion of "robustness analysis", drawing on the earlier work of Campbell (1958), Campbell and Fiske (1959), Levins (1966). Wimsatt also points to Peirce [1868] (1936) and to Whewell's "consilience of inductions" (via Laudan (1971)) as even earlier philosophical precursors. For more recent analyses of the evidential value of robustness, see e.g., Staley (2004), Dethier (2022), Schupbach (2018), Weisberg (2006), Lloyd (2015), and Woodward (2006).

17. See Dethier (2022, 2-3) for a clear discussion of this point.

hypothesis has this property to the extent that the evidence supporting it is varied such that it is independent in the sense described above—i.e., a hypothesis is robust when independent lines of evidence produce convergent results that support it.

Second (and perhaps controversially) I endorse a unified view of robustness that treats agreement across model reports and experimental results in the same way. This is motivated by the idea that models, like measurements, provide a basis for model-based inferences, the results of which can constitute evidence for a hypothesis (Dethier 2022; Tal 2012; Staley 2004, 2020).[18] In the cases I discuss in this paper, the results are robust (or not) across variation in observational methods, messengers, and the physical mechanisms of those messengers' emission. However, robust results can also be achieved by varying models or data analysis methods (see, e.g., Doboszewski and Elder (forthcoming)). I do not think that my main claims in this paper rests on a unified view of robustness (or on a particular unified account of robustness), but I do intend for my account of independent evidence to be able to handle evidence derived from observations/experiments, models, or both.[19]

Independent lines of evidence do not always support consilience arguments. To see why, it is useful to distinguish between "independent lines of evidence" and the arguments made based on these. A single argument that draws on multiple lines of evidence is not consilience if the different lines of evidence do not individually support an argument for the conclusion.

Wylie (2011) distinguishes between "vertical independence" and "horizontal independence". Vertical independence involves different lines of evidence playing different roles in the overall inference. Here, there is a single "linked" argument for the conclusion, where different premises are supported by different pieces of evidence. In linked arguments, there is a "drag-down" effect: the argument as a whole is only as reliable or precise as its weakest premise (Mayo 2018, 15). Such arguments can be represented schematically as follows:

$$E_1 \xrightarrow{\text{supports}} P_1$$

18. See also Schupbach (2018) and Winsberg (2021) for endorsements of the idea of treating robustness across varied experiments and varied models in the same way.

19. For this reason, I eschew terminology referring to narrower or more specific applications of "robustness" that can be found in the literature.

$$E_2 \xrightarrow{\text{supports}} P_2$$

$$(P_1 \& P_2) \xrightarrow{\text{supports}} C$$

Here E_1 and E_2 are independent pieces of evidence and P_1 and P_2 are premises used in an argument supporting the conclusion C. The "supports" arrows here represent some kind of argument in virtue of which the evidence (and later the premises) support a particular claim. This is not a case of consilience because both E_1 and E_2 are needed to support a single argument for C (i.e., neither P_1 nor P_2 implies C—both are needed). However, the Sherlock Holmes strategy could be described by this schematic, since it involves combining multiple pieces of evidence to reach a single conclusion (about a common cause).

In contrast, horizontal independence involves different lines of evidence being used to support separate arguments (ideally, with no shared premises). Here, the inference can be more reliable and precise than any of the individual premises. Mayo (2018, 15) calls this the "lift-off" effect of convergent arguments. The convergent conclusions of these arguments—"consilience"—leads to robust conclusions. At the most basic level, consilience arguments can be represented as follows:

$$E_1 \xrightarrow{\text{supports}} C \qquad\qquad E_2 \xrightarrow{\text{supports}} C$$

$$\therefore C$$

Here E_1 and E_2 are independent pieces of evidence and C is the conclusion that they both support. Whatever confidence each of the pieces of evidence provides for belief in C, the fact that C has these two independent sources of support gives us greater confidence in C than is provided by either E_1 or E_2 alone.

For the convergence itself to be evidence that the conclusion is true, we have to eliminate factors that could lead to convergence on a false conclusion. For example, if the same measuring process (such as a measurement with gravitational wave interferometers) is used to produce both pieces of evidence and this measuring process turns out to be unreliable (e.g., if there was a noise source capable of producing coincident triggers in two separate detectors) then this method could produce convergence without the convergence being evidence for the shared conclusion.

In ideal consilience cases, the lines of evidence are completely independent and hence

the separate arguments for the conclusion C share no premises—they are completely (horizontally) independent of one another. In practice, consilience comes in degrees, according to the level of overlap in premises (where premises correspond to assumptions about auxiliary conditions). The aim of consilience arguments is to improve confidence in the conclusion by showing that *even if* one of the premises in one of the arguments is faulty, the conclusion is not undermined because there is a backup argument that does not rely on the faulty premise. Consilience is an attractive strategy in a trace-based science where many assumptions about the target are not independently testable. This strategy offers the opportunity to make up for weak links in individual arguments where traces are scarce or relatively uninformative about the variables of interest. I discuss two examples of consilience based on GW170817/GRB 170817A/AT 2017gfo below, in sections 4.2 and 4.2.1.

One useful way of gathering independent evidence in astrophysics is using different electromagnetic windows (i.e., different frequency bands) that probe different physical processes in a target system. This entails at least some overlap in the assumptions about auxiliary conditions, given that the different observations rely on at least some shared features of electromagnetic phenomena (e.g., the transmission of electromagnetic radiation through vacuum). However, assumptions about these auxiliary conditions enjoy an extremely high level of empirical support, given the grasp we have on electromagnetic phenomena. Insofar as the shared assumptions in these cases are not likely sources of error, multi-window observations provide the kind of independent evidence needed to support consilience arguments.

Multi-messenger astrophysics looks like an even more promising case for consilience arguments, since there should be even less overlap in the auxiliary conditions required for each messenger to provide evidence about the source. The physical processes of emission, transmission, and detection will all differ. The even greater variation in the auxiliary conditions between different messengers thus leads to greater independence between the lines of evidence and hence, it might be hoped, to strong consilience arguments.

Both of the strategies outlined above—the "Sherlock Holmes" and consilience strategies—rely on scientists gathering a diverse set of traces. However, it is natural to worry that (at least sometimes) the available evidence will be inadequate. For at least some events and processes the recoverable traces may never be sufficient for the task at hand.

Philosophers have disagreed about how serious this problem is, offering a range of

optimistic (e.g., Cleland (2002)) and pessimistic (e.g., Turner (2005)) views. Currie (2018) offers a synthesis of the optimistic and pessimistic views. On the one hand, as time passes, the effects of a past event tend to spread out and diversify. On the other hand, traces also tend to degrade, with the signal becoming faint or getting lost in the noise created by interacting physical processes. These processes of degradation are not generally uniform, but rather tend to differentially preserve certain kinds of traces (e.g., bones) rather than others (e.g., soft tissues). Thus traces both fade and become biased by selection effects. Currie's synthesis captures the insights of both sides, without allowing either to be globally dominant. Instead, the question of whether we have sufficient traces to answer a particular research question will be a local one, depending on the net effects of information-preserving and information-destroying processes at play in the particular context. This seems to me like the right way of looking at things, particularly given the diversity of processes that are relevant across different historical sciences (or even within a single science).

In astrophysics, traces are the result of radiative processes by which messengers carry information over vast distances. These traces are degraded by both distance and by intervening matter. For example, the strength of a gravitational-wave signal decreases with the distance it travels to reach us. In practice, the limited sensitivity of the LIGO and Virgo interferometers means that we can only observe gravitational waves originating from compact binary mergers within a limited volume of space. Since gravitational waves interact very weakly with matter, they are not thought to be significantly affected by the matter they encounter along the way. However, interactions with matter can be both beneficial and detrimental in electromagnetic astrophysics. For example, gravitational lenses allow astrophysicists to learn about massive bodies through their gravitational interactions with electromagnetic radiation originating behind them. On the other hand, the dust layer in the galactic centre renders it opaque to many wavelengths of electromagnetic radiation, which limits the available data about that region. Thus the relatively weaker interactions of gravitational waves with intervening matter can be seen as having both costs and benefits for their usefulness as a messenger.

Even if traces do not seem to significantly diversify over time, as they do in other historical sciences, it is nonetheless possible to obtain diverse traces by tapping into different lines of evidence. Astrophysicists do so by observing through different electromagnetic windows and—where possible—via multiple messengers.

3.3 The Significance of Multi-Messenger Observations

Against this backdrop, there are some clear reasons why multi-messenger observations can be important. These are to do with the multiplication and diversification of the available traces. Increasing the size and diversity of the available trace set increases the dependency relations that can be exploited in order to learn about target systems of interest.

Gravitational and electromagnetic radiation offer independent lines of evidence. Indeed, the range of electromagnetic windows available can offer *multiple* independent lines of evidence. These independent lines of evidence can be used to make inferences about the target system in different ways. One way is captured by the Sherlock Holmes method: the traces can used to infer a common cause that unifies the information provided by each of them. The more diverse the traces, the better the reconstruction of this past event can be. Another way to use independent lines of evidence is consilience: the traces can be used to support independent arguments for the same conclusion. These argumentative strategies offer two different ways of making use of the independence of lines of evidence: *complementary* evidence, where independent lines of evidence are combined to reach a conclusion that none of them could justify in isolation; and *convergent* evidence, where the independent evidence enables consilience-style reasoning (as well as coherence tests). With consilience, redundancy becomes a virtue, increasing the security of an evidence claim. In both cases, having diverse evidence is crucial for getting the arguments off the ground.

Set against the broader concerns about the availability of traces in historical sciences, we see that multi-messenger observations provide extremely lucky circumstances. The independent lines of evidence provided by different messengers allows us to draw conclusions about the particular event, broader astrophysical processes (e.g., the association of binary neutron star mergers with kilonovae and heavy element nucleosynthesis), and even cosmological expansion, all by observing a single target system. I now turn to discussing some of the specific arguments that exemplify different uses of these multi-messenger observations.

4 Arguments from Independent Evidence

4.1 Complementary Evidence

One of the most important conclusions resulting from the multi-messenger observations GW170817/GRB170817A/AT 2017gfo (gravitational waves, gamma rays, and the optical transient, respectively) was the confirmed association of binary neutron star mergers with both sGRBs and kilonovae.[20] To confirm the kilonova hypothesis, astronomers needed to establish both the unambiguous association of the three signals—the gravitational waves, the sGRB, and the optical transient—and that these observations were consistent with a binary neutron star merger and subsequent kilonova.

The association among the three signals was based on a series of common cause arguments, justified by the spatio-temporal coincidence of the signals along with their specific properties—such as 'the evolution of the spectral energy distribution, rapid fading, and emergence of broad spectral features' of AT 2017gfo (Abbott et al. 2017e, 7). In short, the coincidence of the signals could be explained by positing a single past event: a binary neutron star merger and subsequent kilonova.

First, the association between the gravitational-waves, GW170817, and the sGRB, GRB170817, was established based on the spatial and temporal agreement between the two signals. Abbott et al. (2017c, 5-6) report:

> The temporal and spatial p-values are independent quantities, thus the probability that GRB 170817A and GW170817 occurred this close in time and with this level of location agreement by chance is $P_{\text{temporal}} \times P_{\text{spatial}} = (5.0 \times 10^{-6}) \times (0.01) = 5.0 \times 10^{-8}$, corresponding to a Gaussian-equivalent significance of 5.3σ. This unambiguous association confirms that BNS mergers are progenitors of (at least some) SGRBs.

This argument for unifying the detections with a common cause depends on estimates of the background rates of these events; a common cause argument only works when the probability of coincidence of unassociated signals is low.[21] In this case, the rates of the

20. A reminder of some of the technical terminology here: a binary neutron star merger is the collision and coalescence of two neutron stars; sGRB stands for "short gamma-ray burst," a short-lived burst of gamma radiation; a kilonova is a kind of transient astrophysical event featuring ultraviolet, optical, and infrared emission.

21. See Currie (2018, 6) for a good discussion of the importance of reasoning about backgrounds when making common cause arguments in the context of historical sciences.

(detectable, with the given instruments) events are easily computed with some basic assumptions about the background (e.g., isotropy) and measured event rates for both gravitational waves and gamma-rays.

The second step is to demonstrate the further association of the gravitational waves and gamma rays with the optical transient AT 2017gfo. This might initially seem unnecessary—it is tempting to think of the transient as a prediction of the LIGO-Virgo Collaboration that received stunning confirmation once AT 2017gfo was discovered within the region of the sky picked out by GW170817 and GRB170817A. However, this is too quick. Optical transients are relatively frequent events, so once again background rates needed to be accounted for when arguing for the association.

Note that there is a disanalogy between any argument from spatio-temporal agreement made here and the above argument associating GW170817 and GRB170817A. The earlier argument involved two independent measurement processes (by the LIGO-Virgo and *Fermi*) and two independent estimates of the source locations. The agreement itself was thus straightforward evidence for association (given the low background rates). In contrast, the agreement between the optical transient and the other localizations was all but guaranteed by the fact that the search for a transient took place within the region picked out by the gravitational wave (and gamma ray) localizations. Given that the observation campaign was focused on this particular region, it cannot be used to estimate background transient rates in the same way that the LIGO-Virgo or Fermi and INTEGRAL data can. What needs to be determined is the background likelihood of a similar observational campaign finding a transient like AT 2017gfo in a random patch of space of comparable size to the LIGO-Virgo localization—but no such campaign was (or is likely to be) undertaken.

These points about the role of reasoning about backgrounds may seem routine. However, the latter points concerning the association with AT 2017gfo are somewhat obscured in the LIGO-Virgo papers describing this event, which offer no explicit argument for the further association with AT 2017gfo. All that is said is the following:

> The next 24 hr of observation were critical in decreasing the likelihood of a chance coincidence between SSS17a/AT 2017gfo, GW170817, and GRB 170817A. (Abbott et al. 2017e, 7)

This paper also points out that the features of this transient set it apart from other observed transients:

> The optical and near-infrared spectra over these few days provided convincing arguments that this transient was unlike any other discovered in extensive optical wide-field surveys over the past decade (Abbott et al. 2017e, 7)

Combining these two points, the implicit argument for the association seems to be that the background rates of transients like AT 2017gfo must be very low, so a chance coincidence of such a transient with GW170817 and GRB 170817A are small.

This argument is made explicitly by Coulter et al. (2017) and Siebert et al. (2017), who announce and discuss the discovery of the optical transient by the 1M2H team. Siebert et al. (2017) compute estimates based on: (1) transient surveys, such as the long-running Lick Observatory Supernova Search, "LOSS," (Filippenko et al. 2001; Leaman et al. 2011), which place constraints on the rates of various transient events in a given volume of space; (2) the observed properties of AT 2017gfo, including those that differentiate it from previously observed transients; and (3) the non-detection of AT 2017gfo in images of the host galaxy NGC 4993 dating 2.0, 21.0, and 111.8 days prior to its discovery. (3) constrains the time period for which this transient was visible.

The overall observed transient rates from (1) were used to estimate the rate of occurrence of transients within the LIGO-Virgo localization region. Factoring in (2) placed constraints on the maximum fraction of AT 2017gfo-like events out of the total population of transients recorded by the transient surveys.[22] The probability of a chance coincidence is further constrained by the temporal coincidence, based on (3) and on the time of the first detection by 1M2H. Thus Siebert et al. (2017, 5) conclude that '$P_{chance} \leq 9 \times 10^{-6}$ at 90% confidence'. This, combined with non-detection of any other transients within the localization region, provides strong support for the conclusion that AT 2017gfo is associated with GW170817 and GRB170817A.

Finally, having established the association among the three signals, the task of confirming that the optical transient was consistent with a kilonova but not a supernova was accomplished by combining the features of observations across different bands, and how these changed over time. The full range of observations characterizing the transient are summarized and discussed by Abbott et al. (2017e). Recent work by Abelson (2022) reviews the reasoning behind the confirmation of the kilonova model in detail, with

22. Given that none of the transients observed in previous transient surveys were like AT 2017gfo, it is impossible to obtain a precise value for background rates of such events. However, non-detection can be used to place some constraints on the frequency of such events, since sufficiently frequent events *should* have been previously observed in these surveys.

particular emphasis on the need for varied electromagnetic data to tell the whole story: "To remove any piece of that evidence would be to emit a chapter from that story, and thus compromise our ability to predict the ending". Abelson also correctly emphasizes the importance of the evolution of the transient over time as a key prediction of the kilonova model and hence an important aspect of the observational evidence for the model. This mirrors the points made by Abbott et al. (2017e) who conclude that 'the evolution of the spectral energy distribution, rapid fading, and emergence of broad spectral features indicated that the source had physical properties similar to models of kilonovae.' I take Abelson's depiction of the confirmation of the kilonova model to be a case of the Sherlock Holmes strategy, based on a variety of evidence—specifically, complementary evidence that together paints a complete picture of the target system and its evolution.[23]

Overall, the common cause arguments for the association between the signals, combined with the broadband followup campaign, confirmed the association between binary neutron star mergers, sGRBs, and kilonovae. The discovery of the optical transient (guided by the gravitational waves and gamma rays) can be considered the "smoking gun" in this story, which is not to say that the optical transient *alone* constituted persuasive evidence for the conclusion (see above, section 3.2. Detailed analysis of AT 2017gfo combined with GW170817 and GRB 170817A to clinch the case for the kilonova hypothesis. It is worth noting that establishing this hypothesis depended not only on multi-messenger observations of the event itself, but also on the rich evidential corpus of electromagnetic astronomy and astrophysics, such as long-running transient surveys.

23. My interpretation of Abelson on this point is potentially confused by a terminological difference regarding our use of "variety of evidence". Abelson takes "Variety of Evidence" ("VoE") to be an 'epistemic principle': 'Where multiple and heterogeneous types of evidence converge upon model assumptions or outputs, confidence in the representational accuracy of that model is made stronger than if its sources of evidence were homogenous' (Abelson 2022, 133). I don't take the point that Abelson makes about chronological variety of evidence to be about "*converg[ence]* upon model assumptions or outputs" (emphasis mine) at all—rather, the independent evidence is complementary, and supports the conclusion *when taken together*. This doesn't look like an instance of VoE, taken as the epistemic principle characterized by Abelson. However, if variety of evidence here is simply taken to be varied or heterogeneous evidence, then Abelson's depiction of the kilonova case can be seen as one where where varied evidence is combined to support a Sherlock Holmes-style argument.

4.2 Convergent Evidence

It is clear that multi-messenger astrophysics involves the exploitation of multiple, independent lines of evidence. This suggests that at least part of what makes multi-messenger observations "revolutionary" has to do with consilience.

It is natural to hope that consilience could be used to overcome existing worries about the methodology and epistemology of gravitational-wave astrophysics. Since gravitational-wave astrophysics is an emerging field, many aspects of the background theory (e.g., numerical relativity modeling of gravitational waveforms), instrumentation (e.g., ongoing upgrades to LIGO, Virgo, and KAGRA interferometers) and data analysis methods (e.g., search and parameter estimation algorithms) are being developed and refined alongside data collection. In Elder (2023) I provide a more detailed discussion of the interplay between theory, models, and data in this context. Relatedly, I argue that there is a circularity problem when it comes to making inferences about binary black hole mergers on the basis of gravitational waves—that is, "observing" binary black hole mergers. Briefly, the problem is that the empirical validation of models of these systems is based on observations (via gravitational waves) that presuppose the validity of these models.[24] I argue that this problem is compounded by the lack of independent access to the target system by which to break the circularity. Might multi-messenger observations solve this problem? To partially answer this, we can consider the consilience argument in favor of the source of GW170817 being a binary neutron star merger.

The best-measured mass parameter from a gravitational waveform is the chirp mass $\mathcal{M}$, defined as:

$$\mathcal{M} = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}} = \frac{c^3}{G}\left[\frac{5}{96}\pi^{-8/3} f^{-11/3}\dot{f}\right]^{3/5}, \tag{1}$$

where m_1 and m_2 are the individual masses of the components of the binary, and f and $\dot{f}$ are the observed frequency of the gravitational waves and its time derivative. From this, the total mass is estimated to be between 2.73 and 3.29 $M_\odot$, with component masses of 0.86 to 2.26 $M_\odot$ (Abbott et al. 2017d, 2).[25] These values, combined with

24. Note that this is not simply Collins' "experimenter's regress," which concerns the experimenter's confidence that their measuring instruments are operating appropriately and hence able to accurately measure a phenomenon of interest. Rather, this is an extra layer on top of such worries, because complete confidence in the LIGO-Virgo interferometers qua gravitational-wave detectors would not eliminate the epistemic circularity involved in making inferences or "observations" about the source system using models that lack empirical validation aside from those same observations.

25. Parameter estimation by LIGO-Virgo is done within a Bayesian framework, using a stochastic sam-

observed masses of neutron stars and black holes in (other) binary systems, provide strong evidence for the conclusion that the components are neutron stars.

This conclusion is independently corroborated by electromagnetic observations. The discovery of an electromagnetic counterpart rules out a binary black hole merger, since such a merger would not produce an optical transient. The specific properties of the transient, particularly their consistency with models of kilonovae, further support the conclusion that the source was a binary neutron star merger.

Thus we have convergent arguments for the conclusion that the source was indeed a binary neutron star merger. Since these arguments were based on independent gravitational and electromagnetic observations, this is a case of consilience.

This multi-messenger event can be seen as a bridge between the methods of gravitational-wave astrophysics and the more well-established methods of astrophysics as a whole. Both the newer methods of gravitational-wave astrophysics and the entrenched methods of electromagnetic astrophysics can be leveraged to gain empirical traction on this event. Insofar as their results cohere, the methods of gravitational-wave astrophysics may become more trusted and entrenched through the connections thus forged with electromagnetic astrophysics. However, this doesn't do as much as we might hope in assuaging the circularity worries mentioned above.

In particular, this convergence does not address fine-grained concerns about model validation I discussed in Elder (2023). First, the electromagnetic counterparts are not sensitive to the dynamics of the inspiral or the parameters of the source. The photons are emitted in separate physical processes and provide little information about the pre-merger system, except insofar as component masses above a certain threshold would be black holes, producing no electromagnetic counterparts. Second, validating the models used for GW170817 would do little to validate models of compact binary mergers in general, particularly since black holes merge at closer separation distances and more extreme physical regimes than neutron stars do.

However, this convergence does at least suggest that nothing is going terribly wrong with the inferences made based on gravitational waves; it provides an important check on the parameter estimation methods of the LIGO-Virgo Collaboration, insofar as different source parameters would have been incompatible with an observed electromagnetic transient.

Here, it is worth noting a distinction between *consilience* and *coherence testing*

pling library ("LALInference"). See Abbott et al. (2020) and Veitch et al. (2015).

(Bokulich 2020). Both involve using independent methods then checking the consistency of the results. However, they differ in terms of aims. Coherence testing is for validating the methods used to obtain a particular measurement. As Bokulich (2020) argues, discordance can be as useful as concordance, since it indicates the need to take some kind of action (e.g., calibration) that will bring the results into line. On the other hand, consilience is aimed at improving confidence in the results. Here, the idea is to improve the security of an evidence claim despite the possibility of errors in individual methods.

Bokulich argues that coherence testing is prior to calibration and consilience (and indeed, I am convinced by the cases she presents that this is often so). However, this is not true in the case described in above. In the case of GW170817/AT 2017gfo, both kinds of arguments occur simultaneously and are not neatly separable. If anything, consilience is prior to coherence testing (logically speaking). This is because the coherence test only makes sense relative to a judgment about the accuracy of the concordant result. In a case of discordance, we typically think that at least one result is inaccurate, and hence that at least one of the methods must be faulty in some way. But in the case of concordance, we use concordance as a reason to trust the conclusion. Any further judgments about the reliability of the methods used are premised on the accuracy of the concordant conclusion. In the case study of this paper, there is a consilience argument for the conclusion that the event was indeed a binary neutron star merger. Given the concordance (and the empirical entrenchment of the methods of electromagnetic astrophysics) one can *then* make the argument that, because gravitational-wave astrophysics got the correct answer in this case, its methods must also be reliable.

4.2.1 (Non-)Convergent Arguments for the Value of the Hubble Constant

Another potential example of consilience comes from the use of GW170817 to obtain an independent measurement of the Hubble constant (Abbott et al. 2017a). The Hubble constant H_0 is a measure of the rate of expansion of the universe. For distances less than about $50Mpc$, H_0 is well-approximated by the following expression:

$$v_H = H_0 d \tag{2}$$

where v_H is the local "Hubble flow" velocity of the source (the velocity due to cosmic expansion rather than the peculiar velocities between galaxies) and d is the (proper)

distance to the source. To measure the Hubble constant using a given source, one needs to measure both v_H and d for that source. For compact binary mergers d can be measured via gravitational waves, while v_H can be measured from electromagnetic radiation.

Abbott et al. (2017d) reported the distance to the source of GW170817 as $40^{+8}_{-14} Mpc$, based on standard parameter estimation procedures. Abbott et al. (2017a) refines this estimate to $43.8^{+2.9}_{-6.9} Mpc$ by factoring in the location of the optical transient AT 2017gfo in the sky. The ability to use compact binary mergers to measure distances has led to them being dubbed "standard sirens" (analogous to the "standard candles" provided by Type Ia supernovae). This means that compact binary mergers provide a distance measurement that is independent of the "cosmic distance ladder" (and hence independent of the electromagnetic observations used to calibrate this).

The calibration of a cosmic distance ladder has parallels with the calibration of a geological time scale, as discussed by Bokulich (2020). While I leave a detailed examination of these parallels to future work, one parallel that is relevant here is the importance of having independent methods for measuring distance/time that have not been cross-calibrated. Maintaining independence allows both for consilience arguments and for coherence tests where any discordance provides a possible opportunity to refine our methods and learn about systematic errors.

The Hubble flow velocity, v_H, is estimated using redshift measurements. This estimation is based on the identification of NGC4993 as the host galaxy of the transient. The position and redshift of this galaxy are then used to determine the total recessional velocity of AT 2017gfo. Since this event was so nearby (by cosmological standards) the peculiar velocity of the galaxy is expected to be a significant portion of the total recessional velocity. This is corrected for by analyzing the velocities of surrounding galaxies (Abbott et al. 2017a, "methods" section).

Thus using gravitational waves to measure d and the electromagnetic counterpart to measure v_H, Abbott et al. (2017a) obtain a measurement for H_0: $70.0^{+12.0}_{-8.0} kms^{-1}Mpc^{-1}$. This figure comes from a Bayesian analysis and represents the maximum a posteriori value with the minimal 68.3% credible interval. Abbott et al. (2017a) point out that this measurement is broadly consistent with previous measurements of the Hubble constant, despite the independence of the methodology employed: they find 'no evidence [...] for a systematic difference between gravitational-wave-based estimates and established electromagnetic-based estimates' despite the fact that the methods 'may be affected by

different systematic uncertainties' (Abbott et al. 2017a, 86). This is precisely the reasoning behind consilience arguments.

Having a new and independent measurement of H_0 is potentially of great importance, given the notorious *disagreement* between existing measurements. At the time of the GW170817 announcement, the best measurements of this quantity were from Planck (Planck Collaboration et al. 2016) and SHoES ("Supernova H_0 for the Equation of State") (Riess et al. 2016). The Planck Collaboration measurement was based on measurements of the cosmic microwave background (CMB) radiation, including temperature and polarization anisotropies. Given a standard ΛCDM cosmology, this data gives a value of $H_0 = (67.9 \pm 0.9) kms^{-1}Mpc^{-1}$. The SHoES measurement is based on measurements of the distance and redshift of distant galaxies using type Ia supernovae to measure the distance (calibrated using Cepheid variables). Their reported estimate was $H_0 = (73.24 \pm 1.74) kms^{-1}Mpc^{-1}$. The disagreement between Planck and SHoES is often referred to as the "Hubble tension." The discrepancy may be due to systematic errors in one or both measurements, or it may suggest that new physics is needed.

The LIGO-Virgo measurement of $H + 0$ sits in between the *Planck* and SHoES measurements, but with uncertainties large enough to be consistent with both. Given their infamous disagreement, the consistency of the LIGO-Virgo measurement with both can hardly count as a case of convergence or consilience. However, as gravitational-wave astrophysics gains maturity the prospects for such arguments in the future will improve.[26] Future precision measurements of the Hubble constant using standard sirens thus offer a potential source of consilience (and coherence tests).

However, both gravitational wave and electromagnetic measurements are necessary to produce a single measurement by this method. This distinguishes the present case from the previous example of consilience, in which the convergence of independent lines of evidence *about the source of GW170817* was responsible for the consilience. The standard siren case thus exemplifies both complementary and convergent uses of evidence. First, multi-messenger observations of this event are used together to infer the value of the Hubble constant—something that neither the electromagnetic nor the gravitational wave observations could do alone. And second, this overall measurement is independent of other such measurements and provides the opportunity for future

26. Chen, Fishbach, and Holz (2018) estimated that measurements based on standard sirens will be able to constrain the Hubble constant to a precision of 2% within five years (i.e., 2023) and to 1% within the decade.

consilience arguments (and coherence tests).

5 Conclusion

To conclude, I return to the questions that motivated this paper: how do the epistemic resources of multi-messenger astrophysics differ from those of single-messenger astrophysics? And how is this important in meeting the challenges imposed by the epistemic situation of (gravitational-wave) astrophysics? Or, to use more hyperbolic language, what (if anything) makes multi-messenger astrophysics "revolutionary?"

Astrophysics presents us with a difficult epistemic situation. One of the main challenges that characterizes this situation (as for other historical sciences) is the impossibility of performing controlled interventions. Instead, astrophysicists find themselves in a cosmic laboratory, observing traces of the many "experiments" that the universe itself has performed. One of the main tasks of astrophysics is to reconstruct these distant events based on traces. However, in many cases, traces can be insufficient for this task.

A defining feature of *multi-messenger* astrophysics is that the epistemic resources available are expanded to include the methods of multiple domains of observational astrophysics—in this case, the empirically entrenched methods of electromagnetic astrophysics and the developing methods of gravitational-wave astrophysics. Being able to leverage these different methods both expands and diversifies the trace set. This increases the number of exploitable dependency relations between traces and target and enables the kinds of argumentative strategies discussed in this paper—the Sherlock Holmes strategy and consilience. In some cases, the independent lines of evidence buy us novelty—different messengers provide different but complementary information. Here, the independent evidence helps in pursuing a Sherlock Holmes strategy, because adding new kinds of traces improves the chances of clinching the case for a particular hypothesis. In other cases, the redundancy of the information provides the basis for a consilience argument: the convergence of arguments based on independent lines of evidence yields more robust conclusions than the individual arguments alone.

In this paper I have focused on three particular arguments: the argument for the kilonova hypothesis (section 4.1), the argument that the source of GW170817 was indeed a binary neutron star merger (section 4.2), and the "argument" for a new estimate of the Hubble constant, based on an independent, gravitational-wave-based distance

measurement (section 4.2.1). I have suggested that the first case can be analyzed in terms of a series of common cause, or "Sherlock Holmes," arguments. The latter two are cases of consilience. The same multi-messenger observations form the basis for *all* of these arguments. Indeed, some arguments provide a kind of scaffolding on which others can build. For example, the properties of AT 2017gfo are first used to establish the association between the signals. This association is then used as a foundation for inferring the existence of a common cause and confirming that binary neutron stars are progenitors of sGRBs and kilonovae (and hence, via r-process nucleosynthesis, a significant source of heavy elements in the universe). The association is also used in the measurement of the Hubble constant. The consilience arguments can also be turned around to be treated as coherence tests, building confidence in methods that produce concordant results.

The consilience arguments have some clear limitations. One reason for this is the limited overlap in the processes about which the different messengers constitute evidence. The measured gravitational waves (GW170817) are informative about the orbiting binary components, but not very informative about the later kilonova. Likewise, the broadband electromagnetic observations are not sensitive to the details of the pre-merger orbital dynamics. Thus precision coherence tests of the LIGO-Virgo models used for parameter estimation are not possible (and this seems unlikely to change). The second consilience argument, concerning the Hubble constant, offers the possibility of precision tests in the future as more multi-messenger events are observed (Chen, Fishbach, and Holz 2018). However, given that no such events have yet been announced, this remains a promissory note that the LIGO-Virgo Collaboration has yet to cash.

Despite this somewhat negative note, it is important to remember that these multi-messenger observations of a binary neutron star merger put astrophysicists in an extremely fortunate position for reconstructing the event and drawing broader conclusions from it. Due to these lucky circumstances, astrophysicists obtained evidence about astrophysical mechanisms (e.g., neutron stars as a major source of heavy element nucleosynthesis), fundamental physics (via new tests of general relativity) and cosmology (in particular, the Hubble constant). So while this methodological revolution may be largely promissory, these early results are certainly promising. Gravitational waves are a new and independent source of evidence about the universe in a field where scientists are dependent on having diverse sources of evidence. As gravitational-wave astrophysics gains maturity, we will see if multi-messenger astrophysics can live up to this promise.

Acknowledgments

The first draft of this paper was completed during a year as a Heinrich Hertz Fellow at the Lichtenberg Group for History and Philosophy of Physics at the University of Bonn. I would like to thank the Lichtenberg group, especially Dennis Lehmkuhl, Niels Martens, and Juliusz Doboszewski for their support and helpful feedback while producing this paper. I thank Juliusz in particular for letting me sit in on his class "Science without Controlled Experiments" while thinking through the issues discussed in this paper. I would also like to thank the Volkswagen Foundation for its support in providing the funds to create the Lichtenberg Group for History and Philosophy of Physics at the University of Bonn.

I would like to particularly thank Don Howard, Erik Curiel, Nic Teh and Feraz Azhar for their early feedback. Special thanks also goes to Thérèse Arseneau for her careful reading of it during this period.

I also thank audiences at NZAP and BHI Conferences in December 2020, and at HPS-CAP in June 2022 for their insightful questions and comments. Additionally, I would like to thank the attendees of the SuperPAC Workshop in Pittsburgh for starting me thinking about the significance of GW170817 for overcoming the challenges of gravitational-wave astrophysics, and Craig Fox, whose presentations on his own dissertation work encouraged me to think about these issues through the lens of historical science.

Part of this work was undertaken during my employment at Harvard's Black Hole Initiative, which is funded in part by grants from the Gordon and Betty Moore Foundation and the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of these Foundations.

References

Abbott, B. P., et al. 2017a. "A gravitational-wave standard siren measurement of the Hubble constant." *Nature* 551:85–88. https://doi.org/10.1038/nature24471.

———. 2017b. "Estimating the Contribution of Dynamical Ejecta in the Kilonova Associated with GW170817." *The Astrophysical Journal* 850 (2): L39. https://doi.org/10.3847/2041-8213/aa9478.

———. 2017c. "Gravitational Waves and Gamma-Rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A." *The Astrophysical Journal* 848 (2): L13. https://doi.org/10.3847/2041-8213/aa920c.

———. 2017d. "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral." *Physical Review Letters* 119 (16): 161101. https://doi.org/10.1103/PhysRevLett.119.161101.

———. 2017e. "Multi-messenger Observations of a Binary Neutron Star Merger." *The Astrophysical Journal* 848 (2): L12. https://doi.org/10.3847/2041-8213/aa91c9.

———. 2020. "A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals." *Classical and Quantum Gravity* 37 (5): 055002. https://doi.org/10.1088/1361-6382/ab685e.

Abelson, Shannon Sylvie. 2022. "Variety of Evidence in Multimessenger Astronomy." *Studies in History and Philosophy of Science Part A* 94:133–142. https://doi.org/10.1016/j.shpsa.2022.05.006.

Anderl, Sibylle. 2016. "Astronomy and Astrophysics." In *Oxford Handbook of Philosophy of Science,* edited by Paul Humphreys. Oxford University Press, April. https://doi.org/10.1093/oxfordhb/9780199368815.013.45.

Baker, John G, Joan Centrella, Dae-Il Choi, Michael Koppitz, and James van Meter. 2006. "Gravitational-wave extraction from an inspiraling configuration of merging black holes." *Physical Review Letters* 96 (11). https://doi.org/10.1103/PhysRevLett.96.111102.

Bartos, Imre, and Marek Kowalski. 2017. *Multimessenger Astronomy.* 2399-2891. IOP Publishing. https://doi.org/10.1088/978-0-7503-1369-8.

Bokulich, Alisa. 2020. "Calibration, Coherence, and Consilience in Radiometric Measures of Geologic Time." *Philosophy of science* (Chicago) 87 (3): 425–456.

Bokulich, Alisa, and Wendy Parker. 2021. "Data models, representation and adequacy-for-purpose." *European journal for philosophy of science* (Netherlands) 11 (1): 31–31.

Boyd, Nora Mills. 2018a. "Evidence Enriched." *Philosophy of science* 85 (3): 403–421.

———. 2018b. "Scientific Progress at the Boundaries of Experience." PhD diss., University of Pittsburgh.

Boyd, Nora Mills, Siska De Baerdemaeker, Kevin Heng, and Vera Matarese, eds. 2023. *Philosophy of Astrophysics: Stars, Simulations, and the Struggle to Determine What is Out There.* Synthese Library. Springer Verlag. https://doi.org/10.1007/978-3-031-26618-8.

Campanelli, M, C O Lousto, P Marronetti, and Y Zlochower. 2006. "Accurate evolutions of orbiting black-hole binaries without excision." *Physical Review Letters* 96 (11). https://doi.org/10.1103/PhysRevLett.96.111101.

Campbell, Donald T. 1958. "Common fate, similarity, and other indices of the status of aggregates of persons as social entities." *Behavioral Science* (California) 3 (1): 14–25. ISSN: 0005-7940.

Campbell, Donald T, and Donald W Fiske. 1959. "Convergent and discriminant validation by the multitrait-multimethod matrix." *Psychological bulletin* 56 (2): 81–105. ISSN: 0033-2909.

Chen, Hsin-Yu, Maya Fishbach, and Daniel E Holz. 2018. "A two per cent Hubble constant measurement from standard sirens within five years." *Nature* 562 (7728): 545. https://doi.org/10.1038/s41586-018-0606-0.

Cleland, Carol E. 2002. "Methodological and Epistemic Differences between Historical Science and Experimental Science." *Philosophy of Science* 69 (3): 447–451. https://doi.org/10.1086/342455.

———. 2013. "Common cause explanation and the search for a smoking gun." In *Rethinking the fabric of geology,* edited by Victor Baker, vol. 502. GSA special papers. Boulder, Colorado: Geological Society of America.

Coulter, D. A., et al. 2017. "Swope Supernova Survey 2017a (SSS17a), the optical counterpart to a gravitational wave source." *Science* 358 (6370): 1556–1558. https://doi.org/10.1126/science.aap9811.

Currie, Adrian. 2018. *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences.* Life and Mind: Philosophical Issues in Biology and Psychology. Cambridge MA: MIT Press.

Dethier, Corey. 2022. "The Unity of Robustness: Why Agreement Across Model Reports is Just as Valuable as Agreement Among Experiments." *Erkenntnis.*

Doboszewski, Juliusz, and Jamee Elder. Forthcoming. "Robustness and the Event Horizon Telescope: the case of the first image of M87*." *Philosophy of Physics.*

Elder, Jamee. 2020. "The epistemology of gravitational-wave astrophysics." PhD diss., University of Notre Dame.

———. 2023. "Black Hole Coalescence: Observation and Model Validation." In *Working Toward Solutions in Fluid Dynamics and Astrophysics: What the Equations Don't Say,* edited by Lydia Patton and Erik Curiel. SpringerBriefs in History of Science / Technology. https://doi.org/10.1007/978-3-031-25686-8.

Filippenko, Alexei V., et al. 2001. "The Lick Observatory Supernova Search with the Katzman Automatic Imaging Telescope." In *IAU Colloq. 183: Small Telescope Astronomy on Global Scales,* edited by Bohdan Paczynski, Wen-Ping Chen, and Claudia Lemme, 246:121. Astronomical Society of the Pacific Conference Series. January. https://doi.org/10.1017/S0252921100078738.

Forber, Patrick, and Eric Griffith. 2011. "Historical Reconstruction: Gaining Epistemic Access to the Deep Past." *Philosophy and Theory in Biology* 3 (20160405).

Hacking, Ian. 1989. "Extragalactic Reality: The Case of Gravitational Lensing." *Philosophy of Science* 56 (4): 555–581.

Holz, Daniel. 2017. "Hearing and Seeing GW170817," December 5, 2017. Accessed May 30, 2020. https://video.ias.edu/jointastro/2017/1205-DanielHolz.

Jacquart, Melissa. 2020. "Observations, Simulations, and Reasoning in Astrophysics." *Philosophy of Science* 87 (5): 1209–1220. https://doi.org/https://doi.org/10.1086/710544.

Kennefick, Daniel. 2007. *Travelling at the Speed of Thought.* Princeton University Press.

Kleinhans, Maarten G, Chris J. J Buskes, and Henk W de Regt. 2005. "Terra Incognita: Explanation and Reduction in Earth Science." *International Studies in the Philosophy of Science* 19 (3): 289–317. https://doi.org/10.1080/02698590500462356.

Laudan, Larry. 1971. "William Whewell on the Consilience of Inductions." *The Monist* (Chicago) 55 (3): 368–391.

Leaman, Jesse, et al. 2011. "Nearby supernova rates from the Lick Observatory Supernova Search - I. The methods and data base." *Monthly Notices of the Royal Astronomical Society* 412, no. 3 (April): 1419–1440. https://doi.org/10.1111/j.1365-2966.2011.18158.x.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study.* Chicago: University of Chicago Press.

Levins, Richard. 1966. "The Strategy of Model Building in Population Biology." *American scientist* 54 (4): 421–431.

Lloyd, Elisabeth A. 2015. "Model Robustness as a Confirmatory Virtue: The Case of Climate Science." *Studies in History and Philosophy of Science Part A* 49:58–68. https://doi.org/10.1016/j.shpsa.2014.12.002.

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars.* Cambridge University Press. https://doi.org/10.1017/9781107286184.

Mészáros, Péter, Derek B. Fox, Chad Hanna, and Kohta Murase. 2019. "Multi-messenger astrophysics." *Nature reviews physics* (Ithaca) 1 (10): 585–599.

Parker, Wendy S. 2017. "Computer Simulation, Measurement, and Data Assimilation." *British Journal for the Philosophy of Science* 68 (1): 273–304. https://doi.org/10.1093/bjps/axv037.

Patton, Lydia. 2020. "Expanding theory testing in general relativity: LIGO and parametrized theories." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 69:142–153. https://doi.org/10.1016/j.shpsb.2020.01.001.

Peirce, Charles Sanders. 1936. "Some Consequences of Four Incapacities." In *Collected papers of Charles Sanders Peirce,* edited by Charles Hartshorne and Paul Weiss, vol. 5. Originally published in 1868. Cambridge: Harvard University Press.

Pian, E., et al. 2017. "Spectroscopic identification of r-process nucleosynthesis in a double neutron-star merger." *Nature* 551 (7678): 67. https://doi.org/10.1038/nature24298.

Planck Collaboration et al. 2016. "Planck 2015 results - XIII. Cosmological parameters." *A&A* 594:A13. https://doi.org/10.1051/0004-6361/201525830. 10.1051/0004-6361/201525830.

Pretorius, Frans. 2005. "Evolution of binary black-hole spacetimes." *Physical Review Letters* 95 (12). https://doi.org/10.1103/PhysRevLett.95.121101.

Riess, Adam, et al. 2016. "A 2.4% Determination of the Local Value of the Hubble Constant." *The Astrophysical journal* 826 (1): 56.

Schupbach, Jonah N. 2018. "Robustness analysis as explanatory reasoning." *The British Journal for the Philosophy of Science* 69 (1): 275–300.

Shapere, Dudley. 1982. “The Concept of Observation in Science and Philosophy.” *Philosophy of Science* 49 (4). https://doi.org/10.1086/289075.

Siebert, M. R., et al. 2017. “The unprecedented properties of the first electromagnetic counterpart to a gravitational-wave source.” *The Astrophysical Journal Letters* 848 (2): L26. https://doi.org/10.3847/2041-8213/aa905e.

Staley, Kent W. 2004. “Robust Evidence and Secure Evidence Claims.” *Philosophy of Science* 71 (4): 467–488. https://doi.org/10.1086/423748.

———. 2020. “Securing the Empirical Value of Measurement Results.” *The British journal for the philosophy of science* 71 (1): 87–113.

Tal, Eran. 2012. *The Epistemology of Measurement: A Model-Based Account.* http://search.proquest.com/docview/1346194511/.

———. 2013. “Old and New Problems in Philosophy of Measurement.” *Philosophy Compass* 8 (12): 1159–1173. https://doi.org/10.1111/phc3.12089.

———. 2016. “Making Time: A Study in the Epistemology of Measurement.” *The British journal for the philosophy of science* (Oxford) 67 (1): 297–335.

———. 2017. “Calibration: Modelling the measurement process.” *Studies in history and philosophy of science. Part A* (England) 65-66:33–45.

Tucker, Aviezer. 2011. “Historical Science, Over- and Underdetermined: A Study of Darwin’s Inference of Origins.” *The British Journal for the Philosophy of Science* 62 (4): 805–829. https://doi.org/10.1093/bjps/axr012.

Turner, Derek. 2005. “Local Underdetermination in Historical Science.” *Philosophy of Science* 72 (1): 209–230. https://doi.org/10.1086/426851.

van Fraassen, Bas C. 2008. *Scientific Representation : Paradoxes of Perspective.* Oxford; New York: Clarendon Press ; Oxford University Press.

Veitch, J., et al. 2015. "Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library." *Physical Review D* 91 (4): 042003. https://doi.org/10.1103/PhysRevD.91.042003.

Weisberg, Michael. 2006. "Robustness Analysis." *Philosophy of science* (Chicago) 73 (5): 730–742.

Wilson, Katia. 2017. "The Case of the Missing Satellites." *Synthese* 198 (Suppl 21): 1–21. https://doi.org/10.1007/s11229-017-1509-6.

Wimsatt, William C. 1981. "Robustness, Reliability, and Overdetermination." In *Scientific Inquiry and the Social Sciences,* 1st ed., edited by Marilyn Brewer and Barry Collins. San Francisco.

Winsberg, Eric. 2021. "What does robustness teach us in climate science: a re-appraisal" [in eng]. *Synthese (Dordrecht)* (Dordrecht) 198 (Suppl 21): 5099–5122. ISSN: 0039-7857.

Woodward, Jim. 2006. "Some Varieties of Robustness." *Journal of Economic Methodology* 13 (2): 219–240. https://doi.org/10.1080/13501780600733376.

Wylie, Alison. 2011. "Critical distance: Stabilising evidential claims in archaeology." In *Evidence, inference and enquiry.* Proceedings of the British Academy ; 171. Oxford: Published for the British Academy by Oxford University Press.