

Hypothesis Testing and Knowledge of the Human Past


David Černín*

Received: 22 January 2024 / Revised: 24 March 2025 / Accepted: 31 March 2025

Abstract: Historians and natural scientists are adept at inferring knowledge of the past from present traces and evidence. The repository of available methods has been rapidly expanding, and historians of the human past have learned that using techniques developed in other fields that study the natural past might prove beneficial to their endeavours in some cases. The network of inferences involved in historical discourse is vast and diverse. Influential philosophers of history and historical sciences like Jouni-Matti Kuukkanen and Aviezer Tucker have argued that it is necessary to expand the scope of the philosophy of history and to take a deeper look at underlying reasoning and inferential structure of history. This paper answers this call by analysing a group of inferences in human historiography that could be described in terms of hypothesis and hypothesis testing. Hypothesis testing has received some attention in the philosophy of historical sciences, but it is mostly underexplored in the philosophy of history in the context of human history. The paper examines four case studies and analyses their inferential structures by using concepts from the philosophy of historical sciences, such as trace-based and analogous reasoning, type/token distinction, etc. It will be shown that hypothesis testing helps generate knowledge about the past, but

* Ostravská univerzita

 <https://orcid.org/0000-0001-8711-3929>

 Katedra filozofie, Filozofická fakulta, Ostravská univerzita, Reální 5, 701 03 Ostrava, Česká republika

 David.Cernin@osu.cz



to fully appreciate it and to differentiate its types, philosophers of history must engage with the infrastructure of historical research.

Keywords: Hypothesis testing; evidence; experimental archaeology; intellectual history; philosophy of historiography.

1. Outset

Can we arrive at a knowledge of the past via testing hypotheses, and how much can we learn in this manner? In the last few decades, this question has been mostly explored in the realm of historical natural sciences. For example, Derek Turner (2007; 2013) and Carol Cleland (2001, 2013) have engaged in numerous debates regarding the role of prediction and hypothesis testing in geological research. Whereas Cleland argues that *historical sciences* are interested in inferences about token events from the present evidence and they are largely distinct from *experimental sciences*, which operate via using hypothesis and prediction, Turner understands experimental methods and hypothesis testing as a part of a basic toolbox of historical sciences, which can make use of regularities between *type* events (see also Jeffares 2008, Tucker 2011, or Currie 2018).

In contrast, the process of hypothesis testing in human history, historiography, intellectual history, and archaeology remains comparatively unexplored,¹ even though an inquiry into its logic might prove fruitful for our philosophical understanding of historical discourse and its inferential structures. In 1968, Rolf Gruner published a paper explicitly dealing with hypothesis testing in history:

In short, saying that historians test hypotheses when they establish facts is, if not outright false, at least very misleading. It cannot be more than an analogy, but not a good one. A better analogy would be to speak of a pattern of interlocking pieces which

¹ Hypothesis testing receives attention especially in relation to archaeology, see, e.g., “Central Place Theory and the Reciprocity between Theory and Evidence” by Peter Kosso and Cynthia Kosso (1995) or book *Evidential Reasoning in Archaeology* by Robert Chapman and Alison Wylie (2016).

are fitted together similar to the fitting together of a jigsaw puzzle and where the pieces consist of facts of evidence and known historical facts (Gruner 1968, 128).

Much more recently, yet another point of view was offered by the rise of the non-representationalist or postnarrativist philosophy of history. While previous discussions revolved around historical explanations and narratives, the focus shifted towards historical discourse as an argumentative practice. In the words of Jouni-Matti Kuukkanen, the main question is: “How did a historian X arrive at his view of the past given his argumentative context, the sources and the texts available to him?” (Kuukkanen 2017, 118). In 2021, he pursued this project in a text called “Historiographical Knowledge as Claiming Correctly.” Here, he provides a case study of the book *Revolt at Factories* by Seppo Aalto and explores how selected statements from this book are warranted. The background question of this text is whether history is an empirical discipline and what constitutes evidence. Kuukkanen calls for a deeper study of the inferential structures, how historians provide justification for their statements in practice, and how historians decide between legitimate inferential moves and illegitimate ones.

Kuukkanen differentiates between several types of inferences or linguistic acts performed by historians, which he could identify in the book he examined for his case study: “inference from archival material, inference from literature, inference from shared beliefs (historiographical and moral), textual inference, textual coherence, the authority of the historian” (Kuukkanen 2021, 63). Given that this list is founded on one particular case study and one particular piece of historiography, it can not be considered exhaustive, and Kuukkanen does not claim that it is. I do believe that the study of inferential practices that historians employ in their research is a fruitful endeavour that should be pursued in addition to other lines of inquiry in the philosophy of history. However, in this paper, I will claim that we need to expand the scope and depth of our dive into the intricacies of historical inference.

While close reading of historiographic texts may reveal a lot about the logic of historical discourse, it begs several questions. First, if philosophers of history claim to analyse historical discourse, they should be clear about the scope of such enterprise. Should we focus predominantly on extensive

synthesising pieces with clear narrative structure, explicitly worded explanations, and on the reception of these accounts? Or are we to be more inclusive and analyse smaller building blocks of this discourse, like individual papers that focus on establishing the authenticity or relevance of a particular piece of evidential record and focus on minute details in microhistorical studies? What is the role of complex ancillary disciplines like archaeology, aDNA analysis, or experimental ethnographic studies, and how do they factor in the complex network of historiographic discourse and its inferential structures? I would not expect serious objections against a claim that these intellectual activities constitute historical discourse and that we should include written outcomes of these activities in our philosophical reflection of the field.

The second question urges us to go even deeper: Are there inferential activities that are not readily identifiable in every written output of historical discourse regarding a certain topic? Leon J. Goldstein coined a distinction between the infrastructure and the superstructure of history. He defined the superstructure as “that part of the historical enterprise which is visible to nonhistorian consumers of what historians produce” (Goldstein 1976, 141). Thus, the superstructure encompasses literary products (and possibly other types of media) that are the most visible part of historical discourse. Goldstein considers the infrastructure of history to be crucial for philosophical reflection since

it involves treatment of evidence and thinking about evidence and is preoccupied with the determination of what conception of the historical past makes best sense given the character of the evidence in hand (Goldstein 1976, 141).

Even more importantly, Goldstein specifies (1986, 87) that the infrastructure of history concerns activities that are often missing from the final accounts, or they appear only in footnotes, if at all.

Kuukkanen is suspicious of Goldstein’s attempt to draw a thick line between the superstructure and infrastructure:

the distinction between superstructure and infrastructure is not solid, because presentation is a part of the justification of a historiographical work and therefore must be a subject of historiographic epistemology (Kuukkanen 2015, 7).

It could be argued that the superstructure of history exerts a notable power on historians' practice at the infrastructure level. A decision to write a historical account presupposes some idea of the past, knowledge of the discourse and its rules, accepted frameworks, and contemporary terminology. Thus, the preliminary idea about the subject matter may influence the search for evidence and other intellectual activities that Goldstein associates with the infrastructure.

Nevertheless, Goldstein's call for greater attention towards the inferential structures hidden below the surface of textual results of historian's work remains. Interestingly, Goldstein described these intellectual activities in terms of hypothesis testing (1986, 88). The paper will thus focus on these types of intellectual operations in the process of historical discourse and will try to describe their logic.

By a hypothesis in the context of history and archaeology, I mean a statement about the past that receives some degree of support from other theories in the field, but the support is not sufficient to warrant its untested acceptance in historical discourse. In experimental sciences, it is easy to accept Peter Kosso's description: "A hypothesis is a theory that has little testing and is consequently located near the speculation-end of the spectrum" (Kosso 2011, 8). A historical hypothesis is thus viewed as speculation unless additional research activities can somehow decrease the uncertainty, i.e., to support or infirm a hypothesis. "Statements about unobservable things can be tested by their observable implications" (Kosso 2011, 13). Kosso is clear that these implications may include traces of the (unobservable) past, like fossils (2011, 14). In these cases, testing involves searching for additional traces that may serve as evidence, thus increasing consilience. However, we can identify other instances of research activities that may either increase the informational value of available traces and evidence, or produce entirely new evidential bases for the claims about the past, like experimental archaeology or ethnoarchaeology.

This paper will focus on differences between various instances of testing historical hypotheses. The main focus will be on the testing process itself and the bearing of different results on warranted claims about the past, while the inferential activities involved in formulating the hypothesis will be mentioned only where necessary.

2. Evidence-Seeking Hypothesis Testing

Broad historical narratives do not cover the process of historical research in detail. Should we use Goldstein's language, the infrastructure is mostly hidden, except for occasional references to archival material. Authors presuppose (and they are mostly correct) that a general reader is not interested in day-to-day groundwork. In some cases, the context of the discovery of a certain relict or text might be amusing or revolutionary enough to receive a mention in the text or in the footnote. However, if we really are to argue for "practice revolt" (Kuukkanen 2021, 64), then philosophers should search deeper for underlying inferential structures.

The emblematic workspace of a historian is an archive. Historians go there with a certain research question in mind (especially when they need to justify a trip and associated costs according to the guidelines of a grant agency). Be it a study focusing on the life of a particular person or a loftier goal of shedding more light on some pivotal historical event. Archives are a product of peculiar social practice, as noted by Robin G. Collingwood during the dawn of the philosophy of history:

...the past leaves relics of itself, even when these relics are not used by any one as materials for its history; and these relics are of many kinds, and include the relics of historical thought itself, that is, chronicles. We preserve these relics, hoping that in the future they may become what now they are not, namely historical evidence" (Collingwood 1994, 203).

Goldstein, himself being loosely influenced by Collingwood, holds that evidence must always be relative to some previously conceived theory or hypothesis (Goldstein 1962, 180), and Kuukkanen's view of evidence does not seem to stray away, since evidence "can then in general be understood as anything that makes something reasonable to be believed" (Kuukkanen 2021, 64). Yet none of this seems to be at odds with a claim that historians are adept at recognising traces or relicts of the past even before they are identified as evidence for some specific claims or theories about the past, and they participate in practice to preserve, store, and catalogue the traces of the past. In general, it could be argued that traces of the past have some informational value about their origins (Tucker 2025, 4). As historians develop their

methods, they become better at decoding the information contained within the traces, or they learn to recognise additional *types* of traces. These traces may or may not be used as evidence for novel claims about the past. The utilisation of traces and information they contain as evidence is conditioned by research questions that may arise during historical research.

Let us illustrate this process: In the book *Truth and History*, philosopher Murray G. Murphey provides an extensive example of research he was participating in (Murphey 2009, 40–43). The research itself consisted of reconstructing the life of a particular Wyoming physician. Murphey describes this research in painstaking detail and goes over every aspect that baffled historians: sudden changes in the name (Hart/O’Hart) and the date of birth (1890/1891) the historical agent himself reported in archival documents. Murphey focuses on the process of how every discovery provided both some new hypothesis and a clue as to where to look next for another trace – the subsequent potential evidence. It is possible to quote a comparatively brief passage that exhibits all these steps: formulating a likely hypothesis based on available data, identifying ways of testing the hypothesis, testing the hypothesis via further inquiries, and evaluating the results with regard to the original hypothesis.

But we needed to be sure James O’Hart was the father of Hart/O’Hart rather than Patrick. Nebraska did not institute birth certificates until 1904, but if the O’Harts were Catholic, then there should have been a baptismal record. From the Catholic Register for 1890 we found that in 1890 the nearest Catholic Church to Murray was in Plattsmouth—the county seat, and there at the Church of the Holy Spirit we found a baptismal certificate certifying that James Benedict O’Hart had been baptized in June 1890 with parents James O’Hart and Mary Ann Quinn O’Hart and witnesses Patrick O’Hart and Catherine Quinn. So we were now sure that James Oakes Hart was really James Benedict O’Hart, son of James O’Hart, and it was a fair guess that Catherine Quinn was his mother’s sister and probably became Patrick O’Hart’s wife. We were able to confirm that from a newspaper notice of their marriage in December 1890 (Murphey 2009, 42).

To summarise, Murphey states:

At every point, the historian asks himself, if the situation was as I think it was, what should I be able to find, and where should I be able to find it? It should be obvious that the process consisted of a series of predictions and inferences as to what and where the data would be that were then followed up and confirmed or infirmed (Murphey 2009, 43).

In the context of the chapter, Murphey is trying to prove that historians are able to make predictions, i.e., they predict what data they are trying to find and where they will find them; similarly to scientists who are essentially predicting what they will observe under certain conditions (Murphey 2009, 45). However, more interesting is Murphey's description of inferential processes guiding historians from one archive to another, their ability to imagine potential evidence without possessing it and inferring the most probable location of said evidence. This is made possible because historical discourse (as a collection of guidelines and commonly accepted practices) provides historians with some generalised idea of historical evidence.² Furthermore, we have developed a custom of collecting and archiving texts and artefacts that might serve as evidence in the future. Equipped with historical training and awareness of our archival practices regarding documents and artefacts, a historian can propose a testable hypothesis that if x was a case, then there might be evidence y located at $z_1...z_x$.

This (1) *evidence-seeking hypothesis testing* is common throughout historical practice. In some cases, it might direct historians towards previously unconnected parts of contemporary professional historical discourse (e.g., when comparing two well-documented processes that were once considered unrelated). In other cases, it might require archival research (e.g., tracing intellectual influence among medieval scholars requires browsing through unedited original manuscripts and assessing what sources they might have

² The very fact that historians possess some generalised idea of what can be historical evidence is a necessary condition for an insidious practice of creating forgeries and fake artefacts (like the Kensington Runestone) in order to pursue personal or ideological goals. To counter this practice, historians are taught to strictly scrutinise any potential evidence.

had at their disposal, whether they possessed full texts or mere fragments, etc.). (1) *Evidence-seeking hypothesis testing* is an essential part of historians' daily work.

As such, this type of hypothesis is not often explicitly stated in the finished product (e.g., historical narrative), which usually presents a tidy account of past events with a complete list of sources and historical records. Goldstein observed that evidence for various historical claims often appears in footnotes; however, he also acknowledged that even footnotes do not present the entire argumentative background behind the finished account (Goldstein 1986, 87). In the same vein, Kuukkanen states:

In order to study the inferential structure, it is necessary to study the nitty-gritty and follow inferential chains and networks wherever they lead in order to see from where the reasonableness of historians' claiming stem (Kuukkanen 2021, 64).

To pursue these goals, it is imperative to look beyond finished historical accounts and to analyse historians' legwork and their ability to make hypotheses and seek evidence.

We should examine another aspect of the (1) *evidence-seeking hypothesis testing*, which has not been covered extensively by previously mentioned philosophers. When a historian makes a reasonable hypothesis based on available evidence about past events (a historical agent changing the name, intellectual influence between two medieval scholars, social tensions preceding border conflicts), she should be able to postulate possible evidence which would make her hypothesis more probable and its location in order to maintain epistemic diligence. Her subsequent journey to archives or libraries serves as a way to test the hypothesis. Either her research among historical documents will yield results, or she will not find predicted evidence. In a favourable scenario (e.g., a discovery of a letter explaining historical agent's intentions in changing the name; a finding of a copy of a particular historical document authored by an earlier scholar included in a broader unedited manuscript written by a later medieval scholar, which may serve as evidence for intellectual influence) the hypothesis receives some backing.

In a negative case (evidence was not found),³ the hypothesis was neither confirmed nor thoroughly falsified. The hypothesis was “infirm”, as Murphey puts it. The hypothesis about the past might still be plausible or even true, but the evidence was not found and observed, which can mean either that it was not preserved or is located elsewhere. One aspect that Murphey does not discuss in detail is the degree of changes in the probability of a hypothesis given the evidence (or its absence). Suppose we follow Kuukkanen’s call for a deeper study of inferential processes. In that case, we should strive for further examination and learn how exactly different outcomes of this process move a needle regarding the probability of the hypothesis.

Here, significant progress was achieved by philosophers who employ the Bayesian probabilistic framework (Sober 2009; McGrew 2014; Tucker 2025) and discussed how evidence or its absence impacts the probability of a hypothesis. Going back to Murphey’s example, we understand that if we find a baptismal record in the archive, we were able to pinpoint, with the help of the original hypothesis, that it is highly improbable that the record (evidence) would be observed, whilst the hypothesis would be false.⁴ The absence of evidence or – more accurately – our failure to observe the evidence also moves a needle, but its impact is much weaker (Sober 2009, 88–89). McGrew investigated the cases of absence of evidence in relation to historiography, and he noted that if we hypothetically suppose some event to be true, we need to approximate additional probabilities, like how probable it is that somebody would have noticed the event, how probable it is that the event would have been recorded, and how probable is the survival of the given record for a historian to observe (McGrew 2014, 221–222). In the case of the baptismal record as a well-documented established practice, the absence of such a record if the hypothesis is true would seem improbable; thus,

³ In some cases, it is also possible to find data (texts, documents, photos) that directly contradict the original hypothesis. In such cases, the hypothesis should be considered severely infirm and a different claim about the past should be considered strengthened by new evidence. Such a result is more akin to the abovementioned positive case.

⁴ Barring other competing explanations that might be investigated independently, e.g., the baptismal record is a forgery.

the absence of the record will notably decrease the probability of the hypothesis as a whole. However, there are other cases where the absence of a record might not be as decisive. McGrew (2014, 224–225) investigates an interesting case of an absent historiographic record of a major conflagration in Bergen in the early 13th century that is otherwise sufficiently evidenced via the archaeological record. The apparent failure of contemporary chroniclers who did not record the fire is explained via a reference to a theory regarding the record-keeping practices of medieval chroniclers, who often recorded similar disasters selectively and in the context of broader narrative demands (Dunlop and Sigurdsson 1995, 87–89). In other words, the absence of written evidence about the fire is not surprising enough to outweigh the archaeological record and infirm the hypothesis concerning a major conflagration in Bergen because contemporary chroniclers were selective about the fires they recorded. In the more recent past, people may decide not to keep some types of records or to hide them, e.g., because of a fear of persecution by a totalitarian regime (Tucker 2025, 55). Thus, to the result of (1) *evidence-seeking hypothesis testing*, a vast array of theories is necessary, including the theories about the practices of records-keeping in the past.

To summarise, Goldstein urges us to take a look at the infrastructure of historical discourse that is often missing from the final narrative outcomes. With Murphey, he views some of these intellectual practices as a form of hypothesis testing. Following their proposal, we can see that by assessing the impact of evidence (or its absence) on a hypothesis, we uncover another rich network of inferential practices below the simple statements about the past and (in some cases) these practices entail trace-based reasoning and informational epistemology (e.g., Dunlop and Sigurdson 1995 established a theory regarding practices of medieval chroniclers via a sufficiently robust comparative study of numerous traces). Thus, theories concerning transmission and loss of information are crucial in any project that strives to take historical discourse as a serious and rational endeavour practised by a group of experts that is open to warranted revisions. “The general epistemological point is that any claim that can justify (test and verify) what we claim to know about the past must itself be justified” (Kosso and Kosso 1995, 593).

3. Experimental Testing of a Historical Hypothesis

Let us now turn to another discipline, which is, on the one hand, interested in the human past, but, on the other hand, it explicitly frames its methodology in hypothesis testing. There is perhaps no better candidate than *experimental archaeology*. According to Alan K. Outram and his “Introduction to Experimental Archaeology” (Outram 2008, 2), this line of research should be seen as a natural follow-up from archaeological laboratory experiments that inquire into the properties of materials. According to the guidelines proposed by prominent experimental archaeologist Peter Reynolds, a hypothesis is formed on the basis of archaeological data alongside available documentary sources (Reynolds 1999, 127). For example, we may form hypotheses regarding the manufacturing process by analysing ancient pottery and its fragments coupled with historical texts. Such hypotheses can be tested via physical experiments and their replications. In this process, two sets of data are distinguished—original data (i.e., archaeological record and its properties) and data produced by the experiment. Reynolds then states:

If there is agreement between the sets of data, the hypothesis can be tentatively accepted as valid but with the caveat that several different hypotheses raised on the same data might also be validated. If there is no agreement, the hypothesis is not merely invalidated but actually proved to be wrong. The value of this methodology lies especially in the seemingly worst case situation (Reynolds 1999, 127).

This crudely Popperian statement is central to the guidelines of experimental archaeology; however, we may consider it too blunt. Failure to confirm the hypothesis through a well-designed experiment does not lead to invalidation in a straightforward manner (Cleland 2001, 988). It would be more accurate to say that the hypothesis did not receive meaningful support from the experiment. The hypothesis was infirmed and is likely to be rejected by professionals.

An experiment successful at replicating the identified properties of archaeological record shows that this particular approach to manufacturing *might* have been used (leaving space for alternative hypotheses). In contrast,

the failed experiment may substantially infrim the hypothesis, provided that the experiment was designed correctly.

Reynolds himself researched supposed grain-drying ovens (originals dated to the 4th century AD were found in Hertford). He notes that the general consensus among archaeologists about the purpose of these structures goes back to the 1920s, and it used to be largely uncontested—regular findings of traces, such as carbonised seeds, supported this interpretation (Reynolds 1979, 27). Through his experimentation, Reynolds was able to validate a hypothesis that such structures *could have been used* for drying grain (1979, 38). However, in the process of the experiment, he began to doubt whether such drying approach was reasonably efficient or even necessary. “If these structures are not grain driers and in the light of the evidence to date it is improbable, it is necessary to propose an alternative function rather than provide yet another negative” (Reynolds 1979, 41). Given these considerations, Reynolds proposed and later tested an alternative hypothesis – i.e., the structures in question were malting floors. In a later text, Reynolds even provides more background to his reasoning:

A visiting brewer, in the late 1970s, challenged the interpretation of a structure as a Romano-British grain drying oven, suggesting that it was far better as a malting floor. Further research proved his hypothesis to be far more probable! (Reynolds 1999, 134).

(2) *An experimental testing of a historical hypothesis* can be tested through physical experiments. However, as practitioners themselves stress in a Popperian manner, a successful test provides a comparatively weak warrant for a claim about the past. On the other hand, a negative outcome offers a much stronger warrant for a claim regarding the lives of our ancestors. Another aspect of note is that (2) *experimental historical hypothesis* does not help identify any new evidence in the form of past traces like artefacts or texts. It builds upon previously accepted evidence and provides warrants for historical claims about the past through distinct intellectual and experimental activities.

One aspect of (2) *An experimental testing of a historical hypothesis* archaeologists themselves tend to overlook in their methodological musings is a move from *tokens* to *types* and from *trace-based reasoning* to *analogous reasoning*. In some cases, archaeologists are not interested in surviving

artefacts as tokens. They do not inquire into causal chains involved in constructing the particular (supposed) grain dryer; they are not interested in exact years or its builders as individuals. They are interested in the structures as instances of *a type*. Instead of a causal downstream history of the particular artefact, they are interested in the purpose analogous structures could have been regularly used for in the past. The trace-based reasoning (i.e., finding carbonised seeds in the structures) suggested a hypothesis, but experimental testing showed significant issues with such an interpretation (lack of efficiency). The analogous application of expert knowledge from other fields (brewing) helped formulate a competing hypothesis that was not previously available. Despite the insistence of experimental archaeologists that a negative outcome of the experiment is generally more decisive, Reynolds accepted his experiment's positive outcome as a sufficient warrant for labelling the structures in question as malting floors instead of corn dryers. A more robust network of inferences helped to overcome Popperian scepticism about the informational value of a successful experiment.

In this way, experimental archaeology truly works as an experimental science, but we can still understand it as producing a warranted statement about the past. The artificial structures were most likely used as malting floors. This is a knowledge of the past that was not available to archaeologists in the 1920s and led to the mislabelling of actual traces of the past on the basis of crude analogous reasoning. Future traces of the same type that are yet to be discovered will thus inform us about the production capabilities of a particular (token) settlement in a given location. We may hence agree with Ben Jeffares, who argues for blurring the lines between experimental and historical sciences:

The best way to understand the historical sciences is to see them deploying well understood regularities, particular process types, across multiple tokens, either as a means to secure relationships with evidence, or as a general pattern of explanation (Jeffares 2008, 475).

4. Information-Enriching Hypothesis Testing

Well-understood regularities are important for archaeology in general. Let us briefly explore a recent discovery of an artefact with substantial potential for changing established historical discourse. Contemporary historical accounts of Early Slavs in Central Europe are preoccupied with tracing their migration. In the present-day Czech Republic, historical consensus supposes that there was no direct contact between Slavs, arriving after 556 AD, and Germanic tribes (Lombards), who disappeared in approx. 568 AD. The idea that Slavs in the area were not originally influenced by Germanic tribes is a part of accepted historical narratives (e.g., accepted by a community of historians, taught in schools, etc.; see Chlup 2020), and there used to be no evidence to the contrary. However, among the artefacts found at an Early Slavic settlement near Lány, archaeologists have recently discovered a bone fragment bearing a runic inscription (part of *fuþark*). In the context of established historical knowledge, this artefact has significant potential to provide a warrant for a different narrative. Archaeologists, led by Jiří Macháček (Macháček et al. 2021), have recognised this potential due to their professional training, but in order for evidence to move the needle, additional steps were required. To test their hypothesis, they had to employ various techniques to examine the artefact. Through radiocarbon dating, aDNA analysis of animal bones, taxonomical enrichment, runology, etc., the archaeologists have dated the inner bone section to 585–640 AD. Furthermore, they concluded that the inscription was carved by an inexperienced person who was probably practising the runic alphabet; thus supporting the hypothesis that the artefact is evidence for some kind of contact between two ethnolinguistic groups. The exact nature of this contact remains significantly underdetermined (Germanic tribe members living among Slavs, trade exchange, spoils of war, etc.); however, the successful “test of the hypothesis” (i.e., warranted inquiry into the properties of an artefact) increased the strength of evidence, which now warrants a theory going against the accepted historical discourse and can be used as a strong claim in an argumentative historical practice regarding Early Slavic migration in Central Europe.⁵

⁵ It should be noted that professional historical discourse may still raise some objections regarding the finding, e.g., Florin Curta (2009) states that it might be

On the one hand, the testing of the hypothesis required complex technological processes that were facilitated by broad international cooperation. On the other hand, hypothesis testing was straightforward on a theoretical level since it involved (3) *information-enriching hypothesis testing*. Archaeologists do not run the aforementioned laboratory tests without a hypothesis—i.e., without a theoretical suspicion that the artefact may carry some crucial information. If the artefact had been shown to be fake or of a much later date, it would not have been able to function as evidence for a claim concerning the unprecedented contact of two ethnolinguistic groups. The hypothesis would have been adequately infirmed. The successful test of the hypothesis establishes the importance and uniqueness of evidence (i.e., its status as evidence for a certain theory is established), thus achieving a significant milestone in historical discourse. We might say that the theory “is not so much tested as it is used to enhance the informational value of the evidence” (Kosso and Kosso 1995, 591).

Other examples of testing (3) *information-enriching hypothesis testing* may include dating documents, assessing authorship of texts through mathematical modelling, inquiring into chemical properties in the food history context, analysing structural properties of buildings or tools, etc.

Here, we can see that a hypothesis regarding a trace’s informational value *and the potential impact this information can have on a professional discourse* is necessary to justify complex and collaborative testing that enriches the information we can obtain from a particular trace. Without such hope for impactful results, historians and archaeologists do not dedicate their time and resources to every trace they have in their inventory. As such, a number of background theories and inferences are usually hidden from the broader public, but these considerations and the rules of the game of giving reasons play a crucial role in historical discourse. Epistemic diligence dictates that token evidence must undergo significant scrutiny before it can substantiate a claim that undermines previous theories about the past. The information obtained in (3) *information-enriching hypothesis testing* may turn a trace into evidence for a novel and groundbreaking claim about the past, as was the case with runic bone from Lány; or it might

misleading to associate specific material culture with specific linguistic developments.

disqualify a trace as evidence for such claim, e.g., in the case in which the evidence is proven to be of much later date or to be a forgery.

5. Contextualising Historical Hypothesis Testing

Let us now explore one last type of historical hypothesis in this paper. We may encounter explicit hypotheses in another subfield of historical disciplines – namely in intellectual history and related subfields like the history of philosophy. In these cases, historians strive to achieve a better understanding of historical agents and their thoughts or utterances. It may seem that hypotheses and testing have no place in this context. However, when analysing the thoughts of our ancestors, historians often make very explicit hypotheses that serve as tools to contextualise evidence (i.e., text, fragment, etc.). The most explicit use of this approach in a theoretical reflection could be found in the works of Quentin Skinner. Influenced by Collingwood (Skinner 2002, 115), he sees the goal of intellectual history as a recovery of past agent's intentions. To do so without subscribing to Collingwood's idealism, he relies on the speech act theory of John L. Austin (2002, 133). The texts that intellectual historians examine are instances of speech acts with particular illocutionary force (2002, 109). To decode the intention of the past agent, we need to know the context of the speech act. However, the context itself must be established first via historical research, and there are often several contenders for the context historians can utilise while following Skinner's methodological proposition. It might be tempting to say that the context is co-determined by a past actor's intention, but this leads to circularity, as Skinner himself acknowledged in an interview:

I would say that the context is whatever you need to reconstruct in order to understand some meaningful item in that context. This is circular, of course, but I am speaking of a hermeneutic circle. You need to think of texts as answers to questions, and the context as the source of the questions (Li 2016, 122).

Thus, intellectual historians are invited to “test out” different contexts and provide rational argumentation for their choices in order to understand specific utterances. Contexts intellectual historians use are a result of historical

research, and they are often expressed in the form of widely accepted theories within a historical discourse. However, in some cases, intellectual historians might engage in more experimental behaviour.

For instance, exploring ancient Greek philosophy poses multiple challenges to historians. Apart from vast temporal and cultural distance, the often fragmentary nature of texts and gaps in historical records make reaching a consensus about ancient philosophers problematic. A perfect example is Heraclitus of Ephesus' fragmentary and notoriously obfuscated philosophy. Fragments of his work survived through secondary sources, and their reliability and authenticity are often questionable. In his attempt to reconstruct Heraclitus' ethics, David Sider (2013) explores several fragments and tries to contextualise them to achieve a better understanding. A significant portion of this endeavour consists in exploring fragment B29 ("The best choose one thing above all, the everlasting fame of mortals; the many gorge themselves like cattle") as an allusion to Simonides' lyrics dedicated to the fallen at Thermopylae ("Stranger, bring the message to the Spartans that here we remain, obedient to their orders."). The supposed allusion is based on minor linguistic similarities.

Sider acknowledges that the idea itself is a problematic hypothesis in the context of historical discourse regarding Heraclitus because the relative chronology between Heraclitus and Simonides is underdetermined:

The ancient testimony for the death of the former is rather confused, but 484 would seem to be the absolute earliest date; a later date remains quite possible. If so—and this is what I believe to be the case—it seems to more likely that Heraclitus was responding to Simonides (and the favorable reaction his poem no doubt received in Ephesus) than the reverse (Sider 2013, 326–327).

Sider references historical discourse and various critical discussions regarding the year of Heraclitus' death, which makes his hypothesis less probable. At the same time, it is known that Heraclitus often referenced and called out his contemporaries and predecessors, which lends credence to the hypothesis. There is no imaginable method for historians to test this hypothesis unless some supporting evidence is found by pure chance, i.e., this type of hypothesis does not point towards any archive or place where to look. The hypothesis merely contextualises the fragment as a part of a larger

discourse, and its potential interpretative merits are tested only via intellectual reconstruction. In this case, Sider uses the hypothesis to construct an appealing interpretation of Heraclitus' ethics that is consistent with the complex theories of Plato (Sider 2013, 334). Nonetheless, it is questionable whether this "test" (arriving at a plausible and coherent interpretation) provides a solid warrant for the pivotal claim that Heraclitus had been able to allude to Simonides' poem before his death.

In other cases, debates among historians of philosophy may concern, e.g., the relative importance of specific statements in the historical sources as they may contradict other statements in the broader intellectual context of the historical agent's work. A more recent example could be the so-called New Hume debate and the question of whether David Hume was a realist regarding the necessary connexion (see, e.g., Peterková 2015).

We may call it an *interpretative historical hypothesis testing* or (4) *contextualising historical hypothesis testing*. This type of hypothesis is common in subfields like intellectual history, history of political thought, or history of philosophy. It is concerned with interpreting and understanding historical texts, and it is often more explicit than the previous (1) *evidence-seeking hypothesis testing*. It hypothetically holds some statements about the past to be true, connects them to examined texts or thoughts, and the test consists of evaluating resulting interpretations. Even though (4) *contextualising historical hypothesis testing* might be explicitly formulated, the process of testing is not straightforward, and it often consists of historians claiming that their contextualisation provides better results than contesting interpretations in historical discourse. It might be said that we are stretching the meaning of hypothesis testing too far. Yet, these thought operations and related inferences are important in the discourse of intellectual history. The exact nature of this argumentative process might pose a fascinating topic for philosophers of historiography who may study its rules and results.

When (4) *contextualising historical hypothesis testing* allows for meaningful understanding of historical texts (e.g., Sider's daring hypothesis about relative chronology of Heraclitus and Simonides allows for meaningful reconstruction of Heraclitus' ethics), it may be considered as a positive result; however, it does not warrant rejecting other viable interpretations. Contrary to that, when a hypothesis produces incoherent or confusing results, it gains no

support from the process, and it is likely to be rejected by the professional discourse. The evaluation of (4) *contextualising historical hypotheses testing* takes place in an argumentative field of historical discourse.

5. Conclusion

We have explored four types of hypotheses that historians and archaeologists make and how they test them in order to make warranted claims about the past in the context of argumentative historical discourse. This list does not aspire to be exhaustive, and other types of hypothesis testing might be identified further across various historical disciplines.

- (1) *The evidence-seeking hypothesis testing*
- (2) *The experimental historical hypothesis testing*
- (3) *Information-enriching hypothesis testing*
- (4) *The contextualising historical hypothesis testing*

Type (2) is the most typical example of hypothesis testing. The hypothesis must be explicitly worded, and its testing follows strict rules and compares two sets of data. On the surface, its proponents retain strong Popperian principles, and a negative testing result is seen as having a higher informational value than a positive outcome. It relies on *analogous reasoning* and is often concerned with *types*. Hypotheses (3) are needed to explore the evidential potential of historical data that historians or archaeologists identified as promising. Without a hypothesis about the past, complex and expensive laboratory tests are not warranted for every artefact or text. A positive testing of such a hypothesis introduces a new piece of evidence into historical discourse. Negative results of such testing usually do not move the field further.⁶ We do not usually see explicit wording of type (1) hypothesis testing in publications of historians; however, it is an integral part of historians' legwork. The skill to make such hypotheses is crucial for historical practice. By analysing historical discourse and available evidence, a

⁶ However, when a historical record is identified as fake, a different set of historical questions might be asked. Who created the forgery and what was the intention behind it? Exploring these options may produce different hypotheses.

historian hypothesises about the past and potential traces that might be preserved at a specific place (archive, library, etc.). As such, inquiries can be made to test this hypothesis, and the result can be presented in an argumentative context of historical discourse or analysed via Bayesian framework. Finding a relevant document (either confirming or refuting the hypothesis) can be seen as a milestone and may open new avenues for historical discourse. In contrast, failure to obtain relevant data only infirms the hypothesis and does not further historical discourse. Type (4) happens in an argumentative context, does not concern new evidence, and is mostly derived from historical discourse. A hypothesis is often explicitly stated (i.e., it is argued that some context might be relevant, some configuration of past events is presupposed) and tested only via intellectual reconstruction and interpretation. The results of this experimental testing of different contexts are quite surprisingly similar to type (2). A successful (4) *contextualising historical hypothesis* may produce a viable and enriching interpretation, but it may co-exist with other contending interpretations. A failure to produce a coherent interpretation can be seen as a refutation of a hypothesis.

By exploring these selected four types of historical hypotheses, we may see that any excursion into the argumentative context of professional historians as envisioned by Kuukkanen might benefit from going beyond historical narratives and exploring what Goldstein termed the infrastructure of history and which has a long tradition in the philosophy of historical sciences. We may see that the four sketched types of hypotheses testing exhibit different logical structures, different types of reasoning (trace-based, analogous), and some unexpected similarities. Further instances of hypothesis testing can be identified, e.g., Peter Turchin's cliodynamics, which proceeds via testing different hypotheses concerning social and developmental dynamics in history against vast digital databanks of archaeological and historical records (Turchin et al. 2023).

One lesson that should be considered from the present exercise is that if philosophers of history argue for the practice revolt and closer examination of historical discourse and associated argumentative practices, they need to engage with historical discourse on a much deeper level than just on the level of large synthesising pieces of historical literature. Any warranted statement about the past is part of a complex network of inferences that

are difficult to evaluate from reading a single piece of historiography. I do believe that Kuukkanen would not object to this claim. However, so far, it seems that the postnarrativist project is still wedded to the narrativist tenets of exploring historiography as a practice of history writing, and the vantage point does not allow for exploring the infrastructure of history. Philosophers of historical sciences have already shown the potential of exploring the actual practice of historical scientists, and philosophers of history may benefit from closer cooperation. It does not mean that historical narratives and their rhetorical dimensions should be overlooked but that they should not entirely overshadow the underlying infrastructure of history, consisting of evidential, trace-based, and analogous reasoning, informational epistemology, and hypothesis testing.

Funding

This article and research were funded by GAČR: GA23-05800S – “The Big History and its Philosophical Potential.”

Acknowledgements

I am very grateful to the philosophers who have commented on this paper in various stages of production, most notably Aviezer Tucker, Jouni-Matti Kukkanen, Georg Gangl, Eugen Zelenák, and Adam Timmins. I am also indebted to anonymous reviewers for their constructive comments.

References

- Chapman, Robert and Alison Wylie. 2016. *Evidential Reasoning in Archaeology*. London: Bloomsbury Academic.
- Chlup, Radek. 2020. “Competing myths of Czech Identity.” *New Perspectives*, 28: 179–204. <https://doi.org/10.1177/2336825X20911817>
- Cleland, Carol E. 2001. “Historical Science, Experimental Science, and the Scientific Method.” *Geology*, 29: 987–90. [https://doi.org/10.1130/0091-7613\(2001\)029%3C0987:HSESAT%3E2.0.CO;2](https://doi.org/10.1130/0091-7613(2001)029%3C0987:HSESAT%3E2.0.CO;2)
- Cleland, Carol E. 2013. “Common Cause Explanation and the Search for a Smoking Gun.” *Special Paper of the Geological Society of America*, 502: 1–9. [https://doi.org/10.1130/2013.2502\(01\)](https://doi.org/10.1130/2013.2502(01))

- Collingwood, Robin G. 1994. *The Idea of History*, edited by Jan van der Dussen. Oxford: Oxford University Press.
- Currie, Adrian. 2018. *Rock, Bone and Ruin*. Cambridge, Massachusetts: The MIT Press.
- Curta, Florin. 2009. "The Early Slavs in Bohemia and Moravia: A Response to My Critics." *Archeologické Rozhledy*, 61: 725–54.
- Dunlop, Rory and Jón Vidar Sigurdsson. 1995. "An Interdisciplinary Investigation of Bergen's Forgotten Fire: Confrontation and Reconciliation." *Norwegian Archaeological Review*, 28: 73–92. <https://doi.org/10.1080/00293652.1995.9965586>
- Feder, Kenneth. 2019. *Frauds, Myths, and Mysteries: Science and Pseudoscience in Archaeology*. 10th edition. Oxford: Oxford University Press.
- Fritze, Ronald H. 2009. *Invented Knowledge: False History, Fake Science and Pseudo-Religions*. London: Reaktion Books Ltd.
- Goldstein, Leon J. 1976. *Historical Knowing*. Austin and London: University of Texas Press.
- Goldstein, Leon J. 1962. "Evidence and Events in History." *Philosophy of Science*, 29: 175–94.
- Goldstein, Leon J. 1986. "Impediments to Epistemology in the Philosophy of History." *History and Theory*, 25: 82–100.
- Gruner, Rolf. 1968. "Historical Facts and the Testing of Hypotheses." *American Philosophical Quarterly*, 5: 124–29.
- Jeffares, Ben. 2008 "Testing Times: Regularities in the Historical Sciences." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 39: 469–75. <https://doi.org/10.1016/j.shpsc.2008.09.003>
- Kosso, Peter. 2011. *A Summary of Scientific Method*. Dordrecht: Springer. <https://doi.org/10.1007/978-1-4614-1308-0>
- Kosso, Peter and Kosso, Cynthia. 1995. "Central Place Theory and the Reciprocity between Theory and Evidence." *Philosophy of Science*, 62: 581–98.
- Kuukkanen, Jouni-Matti. 2015. *Postnarrativist Philosophy of Historiography*. London: Palgrave Macmillan. <https://doi.org/10.1057/9781137409874>
- Kuukkanen, Jouni-Matti. 2017. "Moving Deeper into Rational Pragmatism." *Journal of the Philosophy of History*, 11: 83–118. <https://doi.org/10.1163/18722636-12341362>
- Kuukkanen, Jouni-Matti. 2021. "Historiographical Knowledge as Claiming Correctly." In *Philosophy of History: Twenty-First-Century Perspectives*, edited by J.-M. Kuukkanen, 44–65. London: Bloomsbury Academic.
- Li, Hansong. 2016. "Ideas in Context: Conversation with Quentin Skinner." *Chicago Journal of History*. 2: 119–28.

- Macháček, Jiří et al. 2021. “Runes from Lány (Czech Republic)—The Oldest Inscription among Slavs: A New Standard for Multidisciplinary Analysis of Runic Bones.” *Journal of Archaeological Science*, 127: 1–8.
<https://doi.org/10.1016/j.jas.2021.105333>
- McGrew, Timothy. 2014. “The Argument from Silence.” *Acta Analytica*, 29: 215–28. <https://doi.org/10.1007/s12136-013-0205-5>
- Murphey, Murray G. 2009. *Truth and History*. New York: State University of New York Press.
- Outram, Alan. 2008. “Introduction to Experimental Archaeology.” *World Archaeology*, 40: 1–6.
- Peterková, Eva. 2015. “David Hume and So-Called ‘New Hume.’” *Perspektywy kultury*, 13: 201–16.
- Reynolds, Peter J. 1999. “Buster Ancient Farm.” In *The Constructed Past: Experimental Archaeology, Education, and the Public*, edited by P. Stone, and Philippe Paniel, 124–35. London: Routledge.
- Reynolds, Peter J. and J. K. Langley. 1979. “Romano-British Corn-Drying Oven: An Experiment.” *Archaeology Journal*, 136: 27–42.
<https://doi.org/doi:10.6067/XCV8JH3Q14>
- Sider, David. 2013. “Heraclitus’ Ethics.” In *Doctrine and Doxography: Studies on Heraclitus and Pythagoras*, edited by D. Sider and Dirk Obbink, 321–34. Berlin: De Gruyter.
- Skinner, Quentin. 1969. “Meaning and Understanding in the History of Ideas.” *History and Theory*, 8: 3–53.
- Skinner, Quentin. 2002. *Visions of Politics*. Volume I: *Regarding Method*. Cambridge: Cambridge University Press.
- Sober, Elliot. 2009. “Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads.” *Philosophical Studies*, 143: 63–90. <https://doi.org/10.1007/s11098-008-9315-0>
- Tucker, Aviezer. 2004. *Our Knowledge of the Past: A Philosophy of Historiography*. Cambridge: Cambridge University Press.
- Tucker, Aviezer. 2011. “Historical Science, Over- and Underdetermined: A Study of Darwin’s Inference of Origins.” *British Journal for the Philosophy of Science*, 62: 805–29. <https://doi.org/10.1093/bjps/axr012>
- Tucker, Aviezer. 2025. *Historiographic Reasoning*. Cambridge: Cambridge University Press.
- Turchin, Peter, et al. 2023. “Explaining the Rise of Moralizing Religions: A Test of Competing Hypotheses Using the Seshat Databank.” *Religion, Brain and Behavior*, 13: 167–94. <https://doi.org/10.1080/2153599X.2022.2065345>

- Turner, Derek. 2007. *Making Prehistory*. Cambridge: Cambridge University Press.
<https://doi.org/10.1017/CBO9780511487385>
- Turner, Derek. 2013. "Historical Geology: Methodology and Metaphysics." *Special Paper of the Geological Society of America* 502: 11–18.
[https://doi.org/10.1130/2013.2502\(02\)](https://doi.org/10.1130/2013.2502(02))