Error Repertoires: Transforming Missteps into Knowledge

Douglas Allchin
Minnesota Center for the Philosophy of Science
University of Minnesota

Abstract. To transform haphazard trial and error into enduring learning, memory is essential. Yet negative results, dead ends and missteps often go undocumented and are rarely shared formally in the journal archive. Still, such "negative knowledge" is often preserved in local contexts and informal communication networks, where it guides subsequent research through ad hoc catalogs of domain-specific missteps, or *error repertoires* (Mayo, 1996). Here, I explore several historical cases and discuss how they function epistemically. Various forms of incongruences — discordant results, theoretical anomalies, and interpretive disagreement — yield new uncertainty and provide researchers a focus for further fruitful research through "troubleshooting," isolating errors, and resolving the apparent discrepancies. However, when investigators find that a former justification was faulty, knowledge grows: a definitive *error* has occurred. Such errors are not useless, residual byproducts. Rather, knowledge of past error can help in interpreting unexpected new experimental results. Identifying possible sources of error also tends to raise evidential standards for subsequent research (at both observational and conceptual levels). Sometimes, they spur new methodologies to forestall or counterbalance errors. Error — and memory of specific sources of error and of general error types — ironically contributes to progress in science.

Introduction: Focusing on Error in Science

Sometimes, experiments go awry. Measuring instruments or observational equipment may require troubleshooting. Results may be unexpected. Findings from different studies may conflict. Researchers may disagree in interpreting the same evidence. Some problems simply stump scientists and remain intractable for decades. Scientific work is unpredictable and challenges the imagination.

On other occasions, however, scientists are outright wrong. What was once considered justified by the evidence falls by the wayside. For example, Galileo was wrong: the tides are not caused by the contrary movements of the Earth. Lavoisier was wrong: heat is not a substance (caloric), and oxygen (despite its name) is not an acidifying principle. On over 150 occasions, claims of discovering new chemical elements proved ill founded (Fontani, Costa & Orna, 2015). These are *errors*: claims once considered fully justified, whose truth status changed with new evidence or revised justification.

In this paper, I discuss how knowledge of past errors can contribute fruitfully to guiding subsequent research. *Negative knowledge* can have a positive role (Allchin, 1999, and forthcoming; Arnaout et al, 2014). First, errors may be characterized in general terms as *error types* and catalogued in a vast *inventory* of error types (Allchin, 2001). These are potential *sources of error* to be avoided in designing subsequent experiments, or monitored or accommodated, if needed. They may also help in diagnosing puzzling or problematic results

when they appear unexpectedly. Analysis of recurrent errors may also inspire the development of new methodological norms, as in the historical origin of controlled experiments or blinded analysis (Allchin, 2012a, 2020, 2023). Yet another way of formalizing the memory of errors addresses the needs of specific fields of study. Scientists are habitually concerned with the particular vulnerabilities of investigative work in their line of research: again, possible sources of error that need to be addressed. We may refer to these local and context-specific catalogs of error types (in contrast to the general inventory) as *error repertoires* (Allchin, 2001, 2019).

The concept of an error repertoire was introduced by Deborah Mayo (1996). “The history of mistakes made in a type of inquiry,” she noted, “gives rise to a list of mistakes to either work to avoid (before-trial planning) or check if committed (after-trial checking)” (p. 5). Namely, in the course of successive investigations by a particular research group, say, or in a particular field, misleading results inevitably arise. As these errors are identified and resolved, one learns what not to do (again). The list of possible errors is an important reference guide.

To date, historians, philosophers and sociologists of science have yet to recognize the importance of Mayo's concept and the ubiquity and significance of error repertoires in scientific practice. Very little seems to have been published documenting the use of error repertoires, or elaborating on the concept — for example, how they are assembled and communicated. This paper begins to address that deficit.

In what follows, I present several historical examples of error repertoires and show how they are integral to ordinary scientific practice and to increasing standards of proof. I then illustrate how they inform the diagnosis of anomalies and other problematic incongruences, in helping to localize a cryptic source of error. Finally, I describe how error repertoires are assembled, highlighting the importance of institutionalized memory.

Escalating Standards

The first, and perhaps most important, dimension of documenting and remembering error is the escalation of standards of proof. Each documented error indicates a plausible path not to pursue, a procedural misstep or confounding factor to avoid, an unreliable proxy variable to jettison, an unwarranted generalization or a misleading theoretical interpretation to set aside. With each documented error, what counts as evidence becomes more stringent. Four examples follow.

Measuring Intelligence

For example, consider the efforts of craniologists to rank intelligence in the 18th-19th century (e.g., Fee, 1979). For many anthropologists and others, measuring the size of the brain—or its surrogate, the skull—seemed to offer an objective way to ascertain mental abilities or capacities. But they encountered a series of errors. Each error led to a change in practice.

At first, the size of the brain or skull seemed an appropriate measure. But this encountered “the elephant problem.” If sheer mass was important, then elephants and whales surely were more intelligent than humans? That seemed intuitively wrong. So, that criterion was abandoned. Size might be important, but not in absolute terms. More discerning evidence was required.

A reasonable alternative seemed to be considering brain size relative to body weight. But that did not yield consistent results either. So, that had to be scrapped too.

Georges Cuvier suggested instead that the size of the cranium relative to the size of the face would reflect a comparatively enlarged brain. Unfortunately, this implied that birds, anteaters and bear-rats would be more intelligent than humans: another counterintuitive conclusion. So: another dead end. For similar reasons cranial height—the relative size of the forehead—was also considered, and then rejected as irrelevant.

Others proposed facial angle, reasoning that an enlarged brain pushed the forehead forwards, creating a more vertical face. That certainly seemed to reflect the evolutionary history of primates. But this ran afoul of apparent exceptions, too, including ranking women and children as more intelligent than men. So, that, too, was added to the growing error repertoire.

Generally, once a measure was shown to be inadequate, no one appealed to that index again (at least to the degree that the knowledge of errors was effectively communicated): the function of an error repertoire. But errors nonetheless continued, resulting in a conspicuous succession of other errors on the very same problem: an *error cascade* (Werth & Allchin, 2021). The error repertoire did not unambiguously specify a solution, but it did raise the bar, thereby limiting the range of eligible claims. By the end of the century, plausible candidates were few indeed.

A more fundamental error was ultimately discovered. Averages had been used to characterize groups, but the group-level differences had been deemed sufficient to compare any two arbitrary individuals between those groups. (Women—as a rule—could thereby be excluded from higher education or political leadership, for instance, regardless of individual variation within each group.) Alice Lee demonstrated the fallacy by publishing the *individual* skull sizes of several prominent male anthropologists alongside those of their *individual* female students. In addition, by capitalizing on the emerging field of statistics, Lee showed that *intra*group differences largely eclipsed by any *inter*group difference. Comparing individuals had no justification. That addition to the error repertoire largely punctured the presumptions of the craniologists' agenda. The whole field began to wane. Boas's discovery in 1910 that the cephalic index was not heritable exposed a further unfounded assumption, contributing further to the erosion of the whole research enterprise. Views relating brain size and intelligence persist, but only among those not fully informed by (or heeding) the historical errors.

Characterizing Human Uniqueness

A similar pattern unfolded in the efforts to characterize humans as unique with respect to tools and tool use (Allchin, 2012b). At first, the use of tools was regarded as distinctive. The earliest hominid fossil associated with tool use was thus honored as marking the emergence of the human genus: *Homo habilis*. But tool use in other species came to light: the woodpecker finch using cactus spines to probe for insects; Egyptian vulture cracking eggs by hurtling stones with their beaks; sea otters cracking open mussels and clams; tailorbird nests; and so on. That applied even to chimpanzees, using twigs to "fish" for ants inside rotting logs.

So, the error was duly acknowledged and registered. No more claims that humans alone used tools. Thinking shifted instead to the *making* of tools (not merely utilizing a nearby artifact). Namely, humans were deemed uniquely *creative*. Well, then we observed chimps making sponges by chewing leaves, and using them to retrieve water from inaccessible spots. They also chewed the ends of twigs to "brush" their teeth. They fashioned pointed spears for killing bushbabies. We now acknowledge that chimps create and leave behind whole tool kits. So: other primates make tools, as well — exposing another error to avoid. The criteria for being

human were raised.

The next claim was that only humans *teach* tool use. Then a chimp was filmed showing an infant how to used a stone to crack a nut against a stationary "anvil" on the ground. At one point, the adult clearly reorients the stone in the infant's hands: unmistakable instruction. Only human? Another error for the error repertoire, not to be repeated. The standards for evidence that established human uniqueness rose again.

The next phase was proposing that only human *prepare* tools for future use. Essentially, nothing to do with tools themselves, but presumably a different telltale indicator of human cognitive abilities. Yet when tested, this claim proved unfounded, as well. The exhibit on human origins at the American Museum of Natural History in New York still characterizes humans in terms of technology, but now in much vaguer terms of degree and kind. The error repertoire has limited what they can justifiably say. Here, the error cascade may illustrate another way this error repertoire may be informative: in indicating an underlying error yet to be fully resolved: the psychological penchant to seek human uniqueness?

Mapping Energy Reactions in the Cell

Error repertoires occur at the experimental level, as well. In the 1950s and 60s, biochemists searched for a series of hypothesized high-energy intermediate molecules in the cell, part of the process known as oxidative phosphorylation, or ox phos (Allchin, 1997). Eventually, 15 claims to have isolated or identified these molecules would be published over 15 years—all of them in error. However, as each artifact was elucidated in turn, ox-phos chemists at least learned from their missteps. Each new error established a new assessment criterion, ostensibly ensuring that chemists would not have to unravel the same error again.

Initially, the claims addressed a short list of properties specified in a 1953 paper that proposed the intermediates (echoed again explicitly in 1966). For example, the compound would form the final energy molecule, ATP (adenosine triphosphate) by reacting with its component parts, ADP and phosphate. Further, the reactions would necessarily be sensitive to a group of chemicals (uncouplers) known to disrupt the process *in vivo*. Those were regarded as the key experimental indicators, based on theory. But as various claims were presented, then rejected, other pitfalls were identified and added to the list of what would be required to demonstrate the molecules experimentally.

By 1963, after eight erroneous claims, Paul Boyer proposed a role for phosphohistidine, published prominently in *Science* magazine. Most of the seven pages of text was devoted to addressing the (by then) many well known sources of error, providing specific evidence to rule each one out. For example, Boyer showed that the energy balance was reasonable, decreasing the likelihood of energy originating from some undocumented source, as was the case for the claim by Jack Purvis in 1958-60. Boyer addressed theories of bonding mechanisms, dispelling potential error in terms of model reactions (what had plagued an earlier proposal by Arnold Brodie). He supplied data on inhibitors and uncouplers. And on exchange reactions and the relative rates of the reactions that they indicated (based on work by Albert Lehninger and Charles Wadkins in 1958-59). That is, he argued his conclusion in part by fully addressing the implicit error repertoire.

Despite heeding the known errors, Boyer's claim, too, proved ill-founded. His results were "real," in a sense, but reflected other reactions in the cell. His method for isolating the ox phos reactions had not materially excluded these other possibilities, which had seemed irrelevant

at the time. So, Boyer's findings resulted in a discovery, just not the one he thought it was originally. His error was evident when, a few years later, he articulated an new explicit list of criteria for an intermediate (an updated error repertoire). Not surprisingly, being able to reintroduce the compound into the cellular extract and produce ATP (while obviously necessary) was not sufficient. Boyer also added other specifics, such as using radioactive labels to trace the purported intermediate through known "exchange reactions." All those criteria raised the standards of proof.

Not all investigators were as thorough as Boyer. For example, when Jui Wang proposed an intermediate in 1971—after 15 proposals had come and gone—critics were quick to note pitfalls that were, by then, widely familiar. Without saying so explicitly, they implied that Wang (or any researcher) was responsible for heeding the error repertoire before presenting conclusions to his peers. Ultimately, presenting data that was simply in accord with theory was not enough. It also had to meet standards of proof based on the implicit error repertoire, shaped by concrete experimental experience.

Determining the Value of Physical Constants

Error repertoires may equally function more locally, by guiding work in a particular lab. For example, consider the classic oil-drop experiment of Robert Millikan for determining the value of the electron's charge (Holton, 1978; Franklin, 1981). Millikan was mindful of many sources of error based on conditions in the lab: such as temperature, convection currents, the role of a fan in distributing the temperature evenly, the position of his floating oil drop relative the edges of his electric plates, and so on (Panusch & Heering, 2011). At first, Millikan's results were widely variable, reflecting the influence of those factors. As Millikan became more familiar with the sources of error in his apparatus, and was able to regulate them, the variability decreased. From October 28, 1911 to February 13, 1912, Millikan observed some 68 drops. None of those appear in his published work. After that date, however, the experiment seemed to be more stable and the calculated values of *e* more uniform. That was when data collection began in earnest: a tribute, in a sense, to exploring — and heeding — his error repertoire.

Successive determinations of many physical constants exhibit a similar pattern. The historical sequence if values converge over time as sources of error are successively encountered and resolved, as unreliable methods are identified, and as corresponding precautionary measures become institutionalized. For example, calculating the rate of the expansion of the universe depended on determining stellar distances. Edwin Hubble had calibrated his work using the special class of celestial objects known as Cepheid variables. Later, Walter Beade discovered that "the" Cepheid variables were composed of two distinct groups, which required separate calibration. The Hubble constant doubled. Several years later Allan Sandage noted some of those "stars" were actually entire nebulae, and that required another recalibration. As each of these errors in the classification of celestial objects was discovered, it set a new standard. And historically, one sees the values begin to converge and stabilize over time (see Huckra, 1992; 2008; graph: Pritychenko, 2015; Trimble, 1996; https://pdg.lbl.gov/2020/reviews/rpp2020-rev-history-plots.pdf; on the Hubble constant, see Pritychenko, 2015).

•

All these cases illustrate how error repertoires are adopted and guide scientific practice. They show how the repertoires are integral to successive claims meeting rising standards, free of

earlier experimental or interpretive errors. In this narrow sense, at least, one may certainly find a degree of progress in science, based on negative knowledge, the cumulative memory of errors.

Knowledge of errors is, ironically perhaps, an asset. For Marvin Minsky (1994), noted pioneer in artificial intelligence, expertise is defined, in part, by never making mistakes. That is, experts know not only how to do things "properly," but *also* how to avoid disasters. They know what *not* to do. Experts are aware of the potential pitfalls in advance. It entails familiarity with exceptions and past mistakes. Experts are, in a sense, masters of the error repertoire in their field (Gartmeier, et al., 2008). Thus, as Minsky observes, that might include scientists "knowing how to intercept and interdict unproductive lines of thought" (p. 18). An expert researcher can spot an impending error even before it happens.

Diagnosing Errors

Error repertoires are not just idle lists of mistakes to avoid repeating, however. They can help to guide fruitful error analysis. For example, when investigators encounter unexpected results — whether inconsistencies among different findings, theoretical anomalous results, or disagreements in conceptual interpretations of the evidence — a reference list of error types can help guide the analysis.

Finding Superconducting Materials

Consider the unfortunate dilemma of post doc Greg Meisner in 1980. He had managed to find a new material with superconducting properties: hafnium rubidium phosphide (HfRbP) (Barz et al., 1980; Stewart et al., 1982). Or so it seemed. In the next stage of investigation, a collaborator used a more sophisticated test to ascertain further properties and establish the alloy's critical temperature more precisely. But the second researcher found no evidence of superconducting. Discordant results. What to do?

Greg's original analysis had used a well established proxy variable: magnetic susceptibility. But this measure (it was well known) was vulnerable to surface shielding, hiding an accurate signal from the bulk interior. Normally, to check for this, Greg would have ground a bit of his material into a powder and measured it again. But the original sample was all gone. He consulted his advisor, whose experience was now critical. In reviewing one of the earlier instrument readings, the advisor noticed a variation in the curve that indicated impurity phases. That is, he knew about and recognized an *error signature*: a distinctive result indicative of a particular kind of error. Verdict?: the compound Greg had measured had been contaminated. Humbled, Greg checked his materials. 99% pure, the manufacturer's label on the hafnium bottle said. Now, it was others in the lab who shared their knowledge of errors. No, they said, do not trust the labels, which could be misleading. Checking with the official specifications in more detail, Greg discovered that, yes, alas, his supposedly "pure" hafnium "may contain 2-3% zirconium." Well, that would easily account for the results, as the superconducting properties of zirconium alloys were already well documented. Sadly, Greg and his colleagues announced the retraction of their results at the next conference, and moved on (Meisner, person. commun.). Knowledge of previous error — including familiarity with a specific *error signature* — had guided them through deciphering the initially deceptive results.

Modeling Climate

This posture of *error analytics* in using error repertoires has inspired philosopher Ryan O'Loughlin (2023) to articulate a strategy for diagnosing climate change models. There are a handful of such models now. They vary in their assumptions and in their core variables, and they yield different, sometimes contrary predictions. The models are recognized as models—and thus admittedly incomplete and not intended to represent reality faithfully in all respects. Still, modelers are interested in diagnosing the errors in their models, toward identifying weaknesses or limitations, which can inform the development of more realistic or powerful models. The complexity of the various models and the diversity of their assumptions and structure makes fruitful comparison difficult. However, a reference list of errors—an error repertoire—may prove a fruitful diagnostic tool.

For example, some models have shown sensitivity to the parameters used to represent clouds: they affect reflection of solar radiation. Others are sensitive to the choice of parameters for convection units — and how they affect precipitation trends. These known sources of errors can be used to assess other models. Background knowledge of physical relationships has also proven useful on occasions. This is a project is yet to be fully realized, but based on the work of some climate modelers, shows promise. Errors known from one context can be used as benchmarks to assess their effect in other contexts.

•

The canonical "Scientific Method" of textbooks is misleading in more than one respect. Most notably, perhaps, it assumes that experiments yield unambiguous results, which allow one to clearly "confirm-or-disconfirm" the hypothesis under consideration. However, more often than we like, results are unexpected—and/or unclear as to their meaning. Troubleshooting experiments and interpreting confusing data is more frequent than philosophers of science may be prepared to admit. Knowledge of error repertoires, however, can be an important tool in diagnosing the puzzling outcomes. The presence of error signatures, in particular, can help investigators isolate cryptic sources of error. Again, negative knowledge can, ironically, have a positive role.

Developing an Error Repertoire

How do error repertoires develop? They are, significantly, a form of memory. They may be assembled by individuals or by particular lab groups, based on local experience (as in the Millikan or superconductivity cases). They may also develop implicitly across a field as errors are identified and percolate through the community (as in the cases of ox phos, craniology, or tool use).

For error repertoires to be effective, they must be accessible through communal memory. That requires publishing negative results and "failures," or communicating them widely through correspondence, hearsay and (more informally, perhaps, but no less inconsequentially) "gossip." Scientific institutions may thus find it appropriate to reflect more on how such negative knowledge is documented and shared, and formalized as part of a searchable scientific archive. Here, I describe the history of two cases of error repertoires, as samples for further discussion.

Diagnosing Brain Death

First, consider the medical problem of determining death. In the late 1960s, conceptions of death shifted toward "brain death": cessation of mental function. Easily said, perhaps. But what were the physically observable medical criteria? An ad hoc committee of experts was convened at Harvard in 1968 to address the challenge of defining an "irreversible coma." Attention focused on three forms of unresponsiveness and the role of an electroencephalograph (EEG) in measuring lack of electrical activity in the brain. But how does one ascertain if the loss of function is permanent, the dimension that essentially signifies death? There were plentiful examples of people who had revived after a coma or loss of consciousness: possible sources of error for a premature or sloppy diagnosis. For the sake of trustworthy medical practice and ethical professional conduct, the criteria needed to be error-free—and transparent. So, the committee included an explicit caveat: repeat the tests again 24 hours later. Also, ensure "that the [EEG] electrodes have been properly applied, that the apparatus is functioning normally, and that the personnel in charge is competent." Those requisite checks, however cursory, were important, of course. The Committee also noted key exceptions, leading to "additional safeguards"— other possible errors to avoid. "The validity of such data," they noted, "...depends on the exclusion of two conditions: hypothermia (temperature below 90 F [32.2 C]) or central nervous system depressants, such as barbituates" (Ad Hoc Committee, 1968, p. 86). Indeed, a subsequent 1977 study by the National Institutes of Health found that a flat EEG, coupled with unresponsiveness, led nearly universally to cardiac arrest. But not completely. Of 187 patients, 2 recovered. Yet those two patients were suffering drug intoxication, and they exhibited reactive pupils. So, that helped confirm the appropriateness of the error repertoire articulated by the ad hoc committee (De Georgia, 2014).

Debates on the philosophical and legal meanings of "brain death" in relation to death continued in the ensuing decades. But the criteria for ascertaining the condition clinically, at least, remained relatively stable. The reliance on an EEG waned, as physicians gained more confidence in clinical means of definitively gauging unresponsiveness and loss of reflexes mediated by the central nervous system. Still, in 2010 the American Academic of Neurology offered further clarity. While they noted that there was "insufficient evidence" to answer some subtle, detailed questions, they were able to establish an explicit checklist for use by practicing physicians. Eight separate brainstem reflexes were specified—including pupils' response to light, gag response, and cough reflex: *all* needed to be checked. They also added such qualifications as: "no evidence of residual paralytics (electrical stimulation if paralytics used)"; "absence of severe acid-base, electrolyte, endocrine abnormality"; "systolic blood pressure $\geq$100 mm Hg"; and "no spontaneous respirations": other conditions where the diagnosis might prove mistaken (Widjicks, et al., 2010). In this way, the error repertoire had evolved further. A further update was issued recently, in late 2023.

Analyzing Neanderthal DNA

As a second case illustrating how error repertoires emerge and deepen, I discuss the history of efforts to recover and sequence Neanderthal DNA, which ultimately led to the 2022 Nobel Prize in Medicine. That outcome was certainly not foreshadowed in the many early failures. The unsuccessful studies, however, helped identify the mishaps and pitfalls that could later be avoided, and thus led to improved isolation protocols and analysis techniques.

Studies began in the early 1990s and benefitted from efforts to recover DNA from other historical specimens. For example, samples of DNA that were extracted from an Egyptian mummy and from a quagga (an extinct relative of zebras and horses) were too small and fragmented to allow anyone else to confirm them through an independent second analysis. So, that indicated an initial limitation. But this was solved by the innovation of PCR (the polymerase chain reaction), which was able to generate many copies from very few.

But, at the same time, this led to the discovery of other problems, exhibited in anomalous and/or discordant results. First, molecular damage could occur (from enzymatic degradation or decay by microorganisms). Short DNA sequences—fragments—could lead to erroneous matching and comparison with modern human DNA. Here, researchers realized that they could integrate the data from different Neanderthal segments by looking for overlaps and reconstructing longer sequences present in the original, even if only a miscellany of eroded fragments remained. One could not avoid this source of error. But it could be accommodated with reconfiguration of the data: a creative methodological solution.

Second, the DNA was susceptible to *copying* errors that occur during PCR itself. To accommodate this "noise," researchers realized that they could compensate with a larger sample. If they performed PCR multiple times on the same DNA sample, they could cross-check the results and identify infidelities amongst the various PCR copies. This enabled them to essentially develop a "consensus" sequence that eliminated the artifacts introduced by their own methods.

Third, contamination by modern DNA was possible—and now plainly detectable. Of course, careful collection and lab hygiene ("clean rooms," protective clothing, sanitized surfaces) had always been a concern. A 1989 paper explicitly listed three additional criteria for confidence in the authenticity of ancient DNA. For example, testing of blank samples was needed as a control, to reveal contamination "introduced from reagents and solutions during the extraction procedure" itself. Also, multiple PCR extractions were again needed — here, to cross-check each other for consistency. As Svante Pääbo reported, these early criteria were "continuously extended as novel aspects of contamination and misincorporations have become obvious." They prompted "the continuous evolution of techniques to avoid contamination." By 2004, the list contained eight criteria—all safeguards against particular errors. For example, a new addition was the criterion of confirmation by a second laboratory (Pääbo, et al., 2004, quotes on p. 655).

In 2006, two independent studies on the genetic divergence of Neanderthals and modern humans were published in the two premier journals, *Science* and *Nature*. Although they were based on the same source of Neanderthal DNA, they did not agree on dating the time of evolutionary divergence: an incongruence. Where was the implicit error? Another team reanalyzed both data sets, and contextualized them in other Neanderthal studies, especially the dating of many fossils (not relying on any DNA analysis). They concluded that the dating in one of the studies was decidedly anomalous, and concluded that the quality of their DNA samples had been compromised (Wall & Kim, 2007). Here, they explicitly drew on the error repertoire to consider the alternatives. Was it (1) contamination with modern human DNA, (2) difficulties aligning Neanderthal DNA fragments, or (3) abnormally high DNA sequencing error rates? By separating the data into sub-groups, they isolated the dating discrepancy to longer DNA segments. That strongly indicated modern human sequences were present and had distorted the overall results. Here, applying an error repertoire was already helping to sort out discordant findings.

Resolving that incongruence certainly did not bring discussion on contamination as a material error to a close. A new method for analyzing the whole Neanderthal genome (shot-gun

sequencing) was introduced, and this brought with it a new set of challenges and potential errors. In addition, as selected elements of the Neanderthal genome were determined with some confidence, they could be leveraged to assess errors in new samples. Measuring congruence with known sequences offered a more reliable baseline for estimating deviation, and thus the likelihood of overall contamination (e..g., Green, et al., 2009). It also provided a basis for, in some cases, editing the results from new samples. Researchers also developed a method of tagging the original Neanderthal sample with a characteristic short nucleotide segment, so that it could be distinguished from any microbial or modern human DNA that might sneak in later in the process (Noonan, 2010). (Think of them as molecular bar codes.) That was another creative innovation. Even if one could not prevent contamination as a source of error, at least one could monitor it and screen it out when it occurred. Again, that involved clear knowledge of the source of error.

Even more remarkably, perhaps, one team developed a method for identifying specific modern DNA sequences in a sample known to be contaminated (Skoglund, et al., 2014). This allowed those sequences to be edited from the results and to yield more reliably authentic sample data. In an initial study, this method allowed a contaminated sample from Siberia to be compared with a Western European specimen, indicating greater convergence than had previously been known. Errors were used to correct other errors.

In summary, an error repertoire for ensuring the reliability (or authenticity) of work on sequencing Neanderthal DNA has grown over three decades. That knowledge guided subsequent work, escalating standards of proof, deepening reliability, allowing quantitative estimates of contamination, and inspiring and informing new methods of DNA sequence reconstruction, and thereby opening opportunities for new discoveries. The history illustrates nicely the many ways scientists leverage negative knowledge to their benefit and ultimately learn from error.

Summary and Prospects

Error repertoires put negative knowledge to active work. They are, ironically, an indicator of expertise. By knowing how to avert disaster, one thereby contributes to a more efficient and reliable research enterprise. Error repertoires can also be important diagnostic tools in troubleshooting incongruences, and in trying to isolate latent, yet-to-be-identified errors. Errors in science may be inevitable. But savvy scientists can catalogue the errors, and learn how to mitigate such interruptions and detours moving ahead, using error repertoires as informative guides in future research. Ultimately, scientists learn from errors.

One frequent product of the elucidation of error repertoires is the development of methods to circumvent those potential errors (Allchin, 2012a; forthcoming). For example, consider Koch's postulates for ascertaining the microbial cause of a disease — not as an a priori set of "postulates," but as an implicit error repertoire, a distillation of historical errors. Koch specified the key ways that one could be mistaken, and thus stipulated the possible sources of error that a responsible investigator needed to address before concluding that any particular bacterium was the cause of a certain disease. Double-blinding doctors in clinical trials, too, has a concrete history, based on growing awareness of the role of observer bias among physicians (Bingel, 1918; Kaptchuk, 1998). In a similar way, one may examine the history of controlled experiments (Allchin, 2020), and see how they emerged as a method to anticipate criticism about alternative explanations: a prophylactic defense against errors. So, too, for many other now-conventional methods, whose origin may be traced historically: the use of placebos/blinding

patients; randomized samples; effective sample size; standardized weights and measures; pre-registration of data analysis protocols and statistical significance cutoffs; and so on. All are implicit error repertoires, transformed into good-sense methodologies (as operational safeguards).

Finally, one may see the "checklist revolution," inspired by Gawande's *The Checklist Manifesto* (2009) as another expression of the use and virtues of error repertoires—and making them explicit. Gawande focused principally on aviation safety and quality control in surgery, where even minor mistakes can be fatal. But he also illustrated how checklists could foster discipline in other fields, including even kitchens at high-end restaurants that serve a severely discerning clientele. In those cases, the costs of error were typically high and the opportunities for recovering from them were quite low. These somewhat extreme conditions helped motivate the rigor of adhering to a formal checklist, and thus avoiding the known repertoire of errors.

Science may not be quite so precarious. But the general spirit of error-checking is similar. In science, I suggest, checklists may perhaps be more informal, but should be supplemented with the epistemic significance of *checkpoints*. That is, at several stages of the research cycle, prior to a major increase in the investment of time and resources, it seems advisable to revisit one's error repertoire— or an inventory of more general error types (Allchin, 2001) — as a way to minimize missteps. Appropriate checkpoints may be at the transition from concept to project design (grant proposal); from plan to action (initiating the material stage of research); from data collection to data analysis (data reduction); from private research to public communication with peers (preparing a paper for submission); and from scientific discourse to relevant social policy (review papers and consensus conferences). These might be the occasions where error repertoires will prove most useful as explicit checklists or reminders of the many possible sources of error.

According to Boumans and Hon (2018), a general philosophy of error is not possible. The prospect of that perhaps dire prophecy remains to be seen. However, as demonstrated in this paper, we can certainly generalize error repertoires as a concept and articulate their functions in general terms. This provides, at the very least, a philosophy of *managing* error. Namely, we can conceptualize the process of learning through error repertoires, and how we can ultimately profit from error by putting negative knowledge to work in our favor.

References

Ad Hoc Committee of the Harvard Medical School to Examine Definition of Death. (1968). A definition of irreversible coma. *JAMA,* 205(6), 85-88.

Allchin, D. (1997). A 20th-century phlogiston: Constructing error and differentiating domains. *Perspectives on Science,* 5, 81-127.

Allchin, D. (1999). Negative results as positive knowledge, and zeroing in on significant problems. *Marine Ecology Progress Series,* 191, 303-305.

Allchin, D. (2001). Error types. *Perspectives on Science,* 9, 38-59.

Allchin, D. (2012a). Teaching the nature of science through scientific error. *Science Education,* 96, 904–926.

Allchin, D. (2012b). To be human. *American Biology Teacher,* 74, 132-135.

Allchin, D. (2019). Error repertoires [presentation]. International Society fo the History, Philosophy and Social Studies of Biology (July 7-12, 2019, Oslo, Norway).

Allchin, D. (2020). The counter-roll in science. *American Biology Teacher,* 82, 188–191.

Allchin, D. (2023). The quixotic quest for objectivity in observation. *American Biology Teacher*, 85, 122-124.

Allchin, D. (forthcoming). *To Err is Science.*

Arnaout, H., Razniewski, S., Weikum, G. & Pan, J.Z. 2021. Negative statements considered useful. *Journal of Web Semantics*, 71, #100661.

Barz, H., Ku, H.C., Meisner, G. P., Fisk, Z. & Matthias, B.T. 1980. Ternary transition metal phosphides: high-temperature superconductors. *Proceedings of the National Academy of Science, USA*, 77, 3132-3134.

Bingel, A. (1918). Über Behandlung der diphtherie mit gewöhnlichem pferdeserum. *Deutsches Archiv für Klinische Medizin*, 125, 284–332.

Boumans , M. & Hon, G. (2018). Introduction. In M. Boumans, G. Hon & A. Petersen (Eds.), *Error and Uncertainty in Scientific Practice*. London: Pickering & Chatto, pp. 1-12.

De Georgia, M.A. (2014). History of brain death as death: 1968 to the present. *Journal of Critical Care*, 29, 673–678.

Fee, E. (1979). Nineteenth-century craniology: The study of the female skull. *Bulletin of the History of Medicine*, 53, 415-433.

Fontani, M., Costa, M. & Orna, M.V. (2015). *The Lost Elements: The Periodic Table's Shadow Side*. Oxford University Press.

Franklin, A. (1981). Millikan's published and unpublished data on oil drops. *Historical Studies in the Physical Sciences*, 11, 185-201.

Gartmeier, M., Bauer, J., Gruber, H. & Heid, H. 2008. Negative knowledge: Understanding professional learning and expertise. Vocations and Learning (2008) 1:87–103 DOI 10.1007/s12186-008-9006-1.

Gawande, A. (2009). *The Checklist Manifesto*. New York: Metropolitan Books.

Green, R.E., Briggs, A.W., Krause, J., Prüfer, K., Burbano, H.A., Siebauer, M., Lachmann, M., & Pääbo, S. (2009). The Neandertal genome and ancient DNA authenticity. *EMBO Journal*. 28, 2494–2502.

Holton, G. (1978). Subelectrons, presuppositions and the Millikan-Ehrenhaft Dispute. *Historical Studies in the Physical Sciences*, 9, 166-224. Reprinted pp.25-83 in *The Scientific Imagination*, Cambridge Univ. Press.

Huckra, J. P. (1992). The Hubble constant. *Science*, 256, 321-325.

Kaptchuk, T.J. (1998). Intentional ignorance: A history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine*. 2(3), 389-433.

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press.

Mayo, D. (2018). *Statistical Inference as Severe Testing*. Cambridge University Press.

Minsky, M. (1994). Negative expertise. *International Journal of Expert Systems*, 7, 13-19.

Noonan, J. P. (2010). Neanderthal genomics and the evolution of modern humans. *Genome Research* 20, 547-553

O'Loughlin, R. (2023). Diagnosing errors in climate model intercomparisons. *European Journal for Philosophy of Science*, 13(2), 1-29.

Panusch, M. & Heering, P. (2011). Robert A. Millikan und die Bestimmung der Elementarladung: Historische Aspekte eines klassischen Experiments. *Naturwissenschaften im Unterricht – Physik*, 126, 32-35.

Pritychenko, B. (2015). A nuclear data approach for the Hubble constant measurements. In *The European Physical Journal Conferences*, p.146.

Skoglund, P., Northoff, B.H., Shunkovc, M.V., Dereviankoc, A.P., Pääbo, S., Krause, J. & Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Science, U.S.*, 111, 2229–2234.

Stewart, G.R., Meisner, G.P. & Ku, H.C. 1982. Specific heats of the new high T_c phosphide superconductors HfRuP and ZrRuP." In W. Buckel, W. Weber (Eds.), *Superconductivity in d- and f-Band Metals: Proceedings of the IV Conference*. Karlsruhe: Kernforshungszentrum Karlsruhe, pp. 331-335.

Wall, J. D., & Kim, S. K. (2007). Inconsistencies in Neanderthal genome sequences. *PLoS Genetics* 3(10), e175.

Werth, A.J. & Allchin, D. (2021). How we think about human nature: Cognitive errors and concrete remedies. *Foundations of Science*, 26, 825–846.

Widjicks, E.F.M., Varelas, P.N., Gronseth, G.S., & Greer, D.M. (2010). Evidence-based guideline update: determining brain death in adults: report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, 74(23), 1911-1918. doi: 10.1212/WNL.0b013e3181e242a8