STAHLp: Belief Revision in Scientific Discovery

Donald Rose and Pat Langley

Department of Information and Computer Science
University of California, Irvine 92717
ArpaNet: drose@ICS.UCI.EDU, langley@ICS.UCI.EDU

Abstract

In this paper we describe the STAHLp system for inferring components of chemical substances - i.e., constructing componential models. STAHLp is a descendant of the STAHL system (Zytkow & Simon, 1986); both use chemical reactions and any known models in order to construct new models. However, STAHLp employs a more unified and effective strategy for preventing, detecting, and recovering from erroneous inferences. This strategy is based partly upon the assumption-based method (de Kleer, 1984) of recording the source beliefs, or premises, which lead to each inferred belief (i.e., reaction or model). STAHL's multiple methods for detecting and recovering from erroneous inferences have been reduced to one method in STAHLp, which can hypothesize faulty premises, revise them, and proceed to construct new models. The hypotheses made during belief revision can be viewed as interpretations from competing theories; how they are chosen thus determines how theories evolve after repeated revisions. We analyze this issue with an example involving the shift from phlogiston to oxygen theory.

I Introduction

Scientific discovery and belief revision are two areas of AI which have undergone considerable investigation, yet work in these areas has rarely overlapped. The STAHL system (Zytkow & Simon, 1986) a forward-chaining production system which constructed componential models of chemical substances was a first step towards combining techniques from both areas. Its domain was 18th century chemistry, during which the prevailing framework was *phlogiston theory*. This theory evolved from the observation that burning substances reduce in size during combustion and thus seem to lose something (phlogiston) in the process. The theory also seemed to explain calcination (now known as oxidization), which was believed to occur when a metal lost phlogiston and transformed into its associated "calx". Thus, phlogiston theory provided rational explanations for two problems which had long frustrated chemists, and indeed seemed to relate both phenomena. In addition to inferring models within this domain, STAHL employed belief revision techniques to resolve conflicts between models, and recover from certain erroneous inferences. However, its methods were limited in scope; we created STAHLp in part to remedy STAHL's deficiencies, but more importantly to further investigate how scientific theories evolve through repeated belief revision.

This work was supported by Contract N00014 84 K 0345 from the Information Sciences Division, Office of Naval Research.

II Overview of STAHLp

Like its predecessor, STAHLp is a forward-chaining production system designed to construct models of chemical substances. Its input consists of 18th century reactions and any known models, and its output consists of newly inferred models. STAHLp's inference cycle begins when premise beliefs are input to the system. Then (1) new models are periodically inferred based on these premises until (2) an erroneous inference is detected. Normal inferencing is then suspended and belief revision begins as (3) hypotheses are generated, proposing ways in which premises would have to be modified to avoid the erroneous inference. Next, (4) the "best" hypothesis having the least impact on existing models is chosen, assigning blame to certain premises; its proposed premise modifications are then carried out. Finally, inferencing (step 1) starts again, (possibly) leading to the construction of more models. Step 1, sufficient if no errors are noted, is itself a cycle in which premises lead to intermediate reactions, then to inferred models, then to more intermediate reactions, and so on.

Like STAHL, the two kinds of beliefs STAHLp deals with are reactions and componential models. Both systems represent a reaction as a list of its *inputs* and *outputs*; a model is represented as a list containing the *substance* being modelled, and its *components*. For example, 18th century chemists observed that calx-of-iron* and charcoal reacted to form iron and ash; STAHLp would represent this reaction as `(reacts inputs {calx-of-iron`

*Like Zytkow and Simon we use 18th century terminology.

charcoal} outputs {iron ash}). We abbreviate this to CI Ch → I Ash. If STAHLp eventually infers that charcoal is composed of phlogiston and ash, STAHLp represents this componential model as (components of {charcoal} are {phlogiston ash}), or Ch = Ph Ash. We can think of these beliefs using an algebraic metaphor; the beliefs above can be viewed as "CI + Ch = I + Ash" and "Ch = Ph + Ash." At this point, we can *substitute* the components of charcoal into the first equation to get "CI + Ph + Ash = I + Ash", then *reduce* ash from both sides to get "CI + Ph = I".

These two steps correspond to the two main rules of both STAHL and STAHLp. Letting S indicate a set of one or more substances, the **SUBSTITUTE** rule is: *If A occurs in a belief, and A is composed of B and S, then replace A with B and S.* The basic **REDUCE** rule is: *If A occurs on both sides of a belief, then remove A from the belief.* When an equation has only one substance in either the inputs or the outputs, STAHLp infers a componential model for that substance, such as I = CI Ph. This third rule, for asserting newly inferred models, is **INFER-COMPONENTS**: *If A and S react to form B, or if B decomposes into A and S, then infer that B is composed of A and S.* At this point, if other reactions are present which contain iron, substitution can occur again; this may in turn lead to further reductions, and more models being inferred, and so on.

STAHLp's basic representation differs from STAHL in that reactions and models are augmented with a *reduced list*. Its main purpose is to keep track, for any belief B, of all substances reduced thus far from reactions which led to the assertion of B. For example, CI Ph Ash → I Ash {~} has an empty reduced list, indicating the REDUCE rule was never applied to reactions leading to it. However, ash can now be reduced, resulting in CI Ph → I {Ash}. At this point, INFER-COMPONENTS would fire, and I = CI Ph {Ash} would be asserted into memory.

III The Belief Revision Process

Zytkow and Simon noted that STAHL has two sources of erroneous inferences: faulty applications of the REDUCE rule; and "error in the input to STAHL" i.e. faulty premises. STAHLp's reduced list plays the key role in handling these problems. In section A we will show its use in *preventing erroneous inferences*. In section B we will show how STAHL's three main error types can be viewed as a single error type in STAHLp, thus allowing a simpler method for *detecting erroneous inferences*. Since REDUCE rule errors have been prevented, and there is only one error type, all erroneous inferences in STAHLp must be caused by faulty premises. In section C we will discuss how such faulty premises are revised, again using the reduced list in this method for *recovering from erroneous inferences*.

A. Preventing Erroneous Inferences

Zytkow and Simon pointed out that "there are situations in which REDUCE produces erroneous conclusions" (Zytkow & Simon, 1986). For example** standard application of STAHL's rules transforms C VA → SA VC and SA = VA Ph into C VA → VA Ph VC after substitution, then into C → Ph VC after reduction of VA. Finally, STAHL asserts a componential model for copper after applying INFER-COMPONENTS: C = Ph VC. However, this conclusion is incorrect; the model of copper accepted in the phlogiston paradigm was C = Ph CC. The missing knowledge needed to construct this correct model of copper is another model, VC = VA CC. If this model had been present as a premise, the correct model would have resulted, because STAHL would have eventually inferred C VA → VA Ph VA CC, then reduced all occurrences of VA. However, if VC's model became known *after* the incorrect copper model was inferred, STAHL would conclude C = Ph VA CC after substitution, which is again incorrect.

STAHL cannot always infer the correct model because it has no mechanism to "remember" what has already been reduced earlier in an inference chain. When the components of VC are substituted into the above reaction, and VA again appears, STAHL cannot remove this occurrence of VA from the reaction and hence cannot infer the correct copper model. However, STAHLp can; it remembers which substances have been reduced through its reduced lists, and removes such substances if they reappear in later (descendent) reactions by using a new rule, **DELAYED-REDUCE**: *If A occurs in a belief, and A also occurs in its reduced list, then remove A from the belief.* For any belief B, this rule ensures that substances previously reduced from B's ancestral beliefs are immediately removed from B. Delayed reduction thus enables STAHLp to construct the correct model C = Ph CC, by removing the new occurrence of VA. In short, the reduced list enables STAHLp to prevent erroneous inferences (e.g. incorrect models like C = Ph VA CC) by revising beliefs as soon as new information (e.g. the components of VC) becomes known.

B. Detecting Erroneous Inferences

Zytkow and Simon classified the erroneous inferences not involving misapplication of the REDUCE rule into three main categories. First, a substance can become defined as being composed of itself (infinite recursion). We refer to this first error type (e.g. where A = B C and B = A D exist concurrently in memory) as a circularity, because a substance assumes a circular definition after applying substitution (e.g. A = A D C or B = B C D). Secondly, there can be two models for the same substance. We refer to this second error type as model subsumption, focusing on the special case where one model's components are a subset of

**C is copper, VA is vitriolic acid, SA is sulfuric acid, VC is vitriol-of-copper, Ph is phlogiston and CC is calx-of-copper.

the other model's components (e.g. A = B C D subsumes A = B C). Finally, a reaction can be inferred where either its inputs or outputs are empty.

We have found that these three error types can be viewed as one type; the first two types can be restated as the third. Thus, a reaction with either empty inputs or empty outputs (but not both) is the fundamental error type in STAHLp; we refer to it as an *unbalanced null reaction* (or simply as an *erroneous reaction*). To see how the first two error types can be restated as unbalanced null reactions, note how the circularity A = B C and B = A D leads to A = A D C after substitution, then to nil = D C after reduction. Conflicting models A = B C and A = B C D lead to B C = B C D after substitution, then to nil = D after two reductions. Thus, the only erroneous inferences STAHLp must detect are unbalanced null reactions, enabling the system to use a simpler, unified method of belief revision.

C. Recovering from Erroneous Inferences

Upon detecting an erroneous inference, STAHLp invokes its main revision process in order to recover from this error. This process decides which premises caused the erroneous inference, revises these premises, and constructs a new theory (i.e. reinfers a new set of beliefs) which does not include the original erroneous inference. In fact, there is historical motivation for this revision method. 18th century chemists sometimes hypothesized missing substances, such as water, in observed reactions, in an attempt to explain conflicting experimental results. For example, Gay-Lussac and Thenard claimed that potassium consists of potash and hydrogen, while Davy observed that potash decomposed into potassium and oxygen. In order to support their claim, Gay-Lussac and Thenard concluded that Davy's potash was not pure but actually contained some water (Zytkow & Simon, 1986). As we shall see later, STAHLp is also able to exhibit such hypothetical reasoning.

STAHLp selects certain premises for revision based on the *source tags* of the detected erroneous inference. Similar to mechanisms in assumption-based systems (de Kleer, 1984), these tags store the underlying premises corresponding to each belief in memory; as a belief B1 is used to infer a new belief B2, B1's tags are propagated to B2. In this way, each belief in memory is associated with the premises that ultimately support it. For each substance in a belief B, its associated source tag contains the substance itself plus the number of *the premise which ultimately contributed that substance* to belief B after a series of rule applications.

1. Generating Effect-Hypotheses

When an unbalanced null reaction is inferred, STAHLp finds the ways in which it could have instead led to a *complete null reaction* (having empty inputs and outputs), by hypothetically altering its LHS or RHS. That is, STAHLp constructs hypotheses about how changes to supporting premises could have the effect of inferring a *balanced* version of this erroneous reaction — one which would lead to a complete null reaction after one or more reductions.

For example, suppose the erroneous reaction is nil → H O {P}. Once detected, STAHLp deletes it from memory and begins belief revision. Its goal is to revise premises such that H and O would not have been left isolated on the RHS. The first step is to perform an "inverse reduction" of all reduced list substances, plugging them back into the reaction; the result here is P → P H O {~}, the *modified erroneous reaction*. Now the system finds how many ways it can change this reaction (without using new substances) so that its LHS equals its RHS. There are four options here: (1) Add H O to the LHS, (2) Add H to the LHS and Delete O from the RHS, (3) Add O to the LHS and Delete H from the RHS, (4) Delete H O from the RHS. These are STAHLp's *effect-hypotheses* — changes to the modified erroneous reaction that would have resulted if certain premises had been different. For example, the balanced reaction P H → P H would be inferred instead of P → P H O if the effect of revising premises is hypothesis (2).

2. Generating Cause-Hypotheses

The problem now is to decide which premises should be revised, by matching each substance in the effect-hypotheses to a corresponding substance in some premise. In the case of substances which must be hypothesized as really absent from some premise (i.e. when the desired effect is deletion from the modified erroneous reaction), there is little complication; the source tag for each "Delete" substance in an effect-hypothesis indicates which premise is involved (as well as which side). For example, given effect-hypothesis (4) and the modified erroneous reaction P → P H O, a tag (H 2 r) would indicate that the RHS of premise 2 should not have had H, while (O 3 r) would indicate that the RHS of premise 3 should not have had O. Thus, STAHLp would construct *cause-hypothesis* (4), whose proposed revisions would result in effect-hypothesis (4): *The RHS of premise 2 must not have had H, and the RHS of premise 3 must not have had O.*

While such commission cause-hypotheses are relatively easy to create, omission cause-hypotheses are more difficult. For each "Add" substance in an effect-hypothesis, STAHLp must decide which premise this substance should have been present in to get the desired effect. The problem is that there is no obvious source tag to work from, since STAHLp will add this substance to a premise it did not exist in before. Our solution is to use the source tags of substances that were plugged back to the *empty* side of the erroneous reaction — substances that are now on the "smaller" side of the modified erroneous reaction. This is the side where substances must be added to effect a balanced reaction. Again using the above example, STAHLp

would create the following hypothesis as the cause of effect-hypothesis (1): *the P on the LHS must really have contained H and O.*

The question now is *which* LHS must be revised — i.e. which premise's P is to blame. P's source tag holds the answer. If it was (P 3 1), the detailed cause-hypothesis (1) would result: *the P on the LHS of belief 3 really had H and O.* (Note that such a conclusion, which STAHLp indeed made in one of its runs, models Gay-Lussac and Thenard's claim that the potash (P) in the reaction Davy allegedly observed really had some water (H and O) in it.) STAHLp simply used P to pinpoint the relevant premise and side in which to hypothesize omitted substances. In general, all reduced-list substances that are plugged back into the empty side of an erroneous reaction are used for this purpose — to aid in constructing omission cause-hypotheses. While we have shown how STAHLp would form cause-hypotheses (1) and (4), the reasoning described would also be used to construct hybrid cause-hypotheses — those containing both omission *and* commission errors (e.g. cause-hypotheses (2) and (3)).

3. Choosing a Best Hypothesis

Having generated detailed cause-hypotheses about how to revise premises in order to avoid the erroneous inference, STAHLp now selects which of these sets of revisions to execute. This step begins when STAHLp computes the *cost* of making the modifications suggested by each cause-hypothesis. The cost reflects how many existing models are supported by those premises which would be changed if a certain cause-hypothesis is applied. For example, let us examine cause-hypothesis (4): *The RHS of premise 2 must not have had H, and the RHS of premise 3 must not have had O.* If premise 2 supports 7 models and premise 3 supports 1 model, then the total cost (of making modifications to premises 2 and 3) is 7 + 1 = 8. After computing the cost of each cause-hypothesis, STAHLp selects the one with the lowest cost — i.e., the one whose revisions will have the least impact on existing beliefs — as the *best hypothesis.*

4. Constructing a New Theory

After choosing the best hypothesis, STAHLp starts constructing a new theory containing a possibly different set of componential models; some may be new, some may no longer present, and others may have been modified. First, for each premise that will be changed due to the chosen hypothesis, all beliefs based on that premise are deleted from memory.[*] Second, the system performs the changes proposed in the hypothesis and asserts the modified premise(s) into memory. Finally, any new inferencing that may occur in response to the new premise(s) is performed. Hopefully, the result will be new models, but at the very least the original erroneous reaction will not be reinferred; the design of STAHLp's revision strategy guarantees that a complete null reaction will result instead during the new inferencing cycle. By viewing the result of reinferencing as the construction of a new theory (i.e. a new set of componential models), one can visualize an initial theory incrementally evolving in response to repeated detections of erroneous reactions and subsequent revisions of selected premises.

[*] To do this, STAHLp looks at the source tags of each belief. If substances from the premise to be changed ultimately contributed to this belief, then at least one of this belief's tags must also be one of the premise's tags, and hence this belief would be deleted.

IV Phlogiston vs. Oxygen

Thus far we have discussed the fundamentals of STAHLp's operation. Let us now synthesize the previous sections by walking through a detailed example, beginning with the assertion of three premises:

```
(1 (M --> CM Ph {~}))
(2 (CM = M O {~}))
(3 (M CI --> I CM {~}))
```

These premises then lead to two inference chains (4 through 6, then 7 and 8):

```
(4 (M CI --> I M O {~}))  after substituting 2 into 3,
(5 (CI --> I O {M}))      after reducing 4,
(6 (CI = I O {M}))        after infer components from 5

(7 (M --> M O Ph {~}))    after substituting 2 into 1,
(8 (nil --> O Ph {M}))    after reducing 7.
```

At this point, reaction 8 (an erroneous inference) is removed and belief revision begins. STAHLp starts generating hypotheses about how the erroneous reaction could have been avoided — i.e. how the substances in the non-empty side of reaction 8 could have been reduced themselves, thus preventing an unbalanced null reaction from being asserted. The answer comes by recognizing the different ways in which a complete null reaction would have resulted instead. The system first constructs the modified erroneous reaction by plugging M back into both sides — in effect turning its attention to reaction 7. Then, STAHLp analyzes the four balanced reactions which might have been inferred instead of reaction 7 if premises had been different (without using any new substances):[†]

```
(EH1) M [O Ph] --> M O Ph   needed O and Ph on LHS;
(EH2) M [O] --> M O (Ph)    needed O on LHS, no RHS Ph;
(EH3) M [Ph] --> M (O) Ph   needed Ph on LHS, no RHS O;
(EH4) M --> M (O) (Ph)      needed no RHS O, no RHS Ph.
```

[†] In the following figure, [] contains substances which would need to be present to produce a balanced reaction, and () contains substances which would need to be absent.

To determine which premises contributed each substance in reaction 7, STAHLp must analyze it complete with its source tag information: (M 1 l) → (M 2 r) (O 2 r) (Ph 1 r) {~}. Now the corresponding cause-hypotheses can be generated. The tag (M 1 l) is used for hypothesizing new LHS substances (omission errors), while the tags for O and Ph are used for hypothesizing commission errors:

```
(CH1) Belief 1, LHS: should have had O and Ph;
(CH2) Belief 1, LHS: should have had O,
      Belief 1, RHS: should not have had Ph;
(CH3) Belief 1, LHS: should have had Ph,
      Belief 2, RHS: should not have had O;
(CH4) Belief 1, RHS: should not have had Ph,
      Belief 2, RHS: should not have had O.
```

Now the cost of carrying out the changes these cause-hypotheses recommend is computed.[‡] Belief 2 supports one model (CI = I O), and belief 1 supports none. Thus CH3 and CH4 have cost 1, since both propose changing beliefs 1 and 2, while CH1 and CH2 have zero (and hence the lowest) cost, since both propose changing belief 1. Let us say CH2 is arbitrarily chosen as the best hypothesis. At this point, all beliefs based on belief 1 (the premise to be modified) would be deleted; here the only belief supported by belief 1 is the erroneous reaction, which has already been deleted. STAHLp now asserts belief 9, a modified version of premise belief 1 which incorporates the changes of CH2: M O → CM {~}. Now reinferencing begins; the substitution of belief 2 into 9 leads to belief 10: M O → M O {~}. Two reductions then lead to a complete null reaction, which is harmlessly deleted from memory. Thus, while no new models were discovered upon reinferencing here, the complete null reaction that resulted shows that if some oxygen was actually present in the inputs of reaction 1, and phlogiston was actually not in the outputs, no beliefs contradicting existing models will be inferred.

STAHLp's hypotheses loosely model how followers of one paradigm can propose revisions of data reportedly observed by followers of another paradigm. For example, a follower of Lavoisier would probably be prone to believe hypothesis CH2, since he would believe in the existence of oxygen, but not phlogiston. The important point of this example, although the number of beliefs was kept small, was that after CH2's revisions were executed *phlogiston no longer existed in any beliefs.* If one defines oxygen theory as a system of reactions and models which do not include the existence of phlogiston, then over a period of time it is possible for STAHLp to revise its set of beliefs from one embodying phlogiston theory to one embodying oxygen theory. Such a theory shift is not guaranteed, but the hypotheses at the very least represent the views of the competing theories; in the best case, the repeated presence of erroneous reactions would lead to removal of phlogiston from premises and hence force the theory shift to take place.

The revision in this example corresponds to how a believer in oxygen theory could analyze beliefs from another paradigm (i.e. phlogiston theory), and hypothesize that those beliefs were actually misinterpreted observations. Similar results were obtained in modelling the dispute between Davy and Gay-Lussac/Thenard. STAHLp revised Davy's premise P → K O to include H and O on the LHS replicating Gay-Lussac and Thenard's reinterpretation of Davy's results. In another example, given a set of 5 Lavoisier-era reactions, STAHLp's belief revision process led to the hypothesis that caloric does not exist – a belief eventually accepted by chemists just as phlogiston's nonexistence was.[††] In short, STAHLp's mechanism for revising premises in response to erroneous inferences enables it to question its basic assumptions, as well as propose new ones – a vital ability in any domain of scientific discovery.

V Summary

STAHLp, a system for constructing componential models of chemical substances, employs a more unified and effective strategy for dealing with erroneous inferences than its predecessor STAHL. The reduced list contains information needed for preventing erroneous inferences caused by misapplied reduction. Detecting erroneous inferences is simpler in STAHLp; STAHL's 3 main error types can be viewed as unbalanced null reactions. Finally, the reduced list enables STAHLp to recover from such an erroneous reaction, using information about where its substances came from to propose revisions to some of its premises. Once a plausible hypothesis is chosen to account for the error, the premises it assigns blame to are revised, and a new set of beliefs are eventually inferred. STAHLp's main contribution lies in its incorporation of more powerful belief revision techniques into work on scientific discovery, and in its potential for modelling how theories evolve.

References

de Kleer, J. Choices without backtracking. *Proceedings of the Fourth National Conference on Artificial Intelligence*, Austin, Texas (1984) 79–85.

Zytkow, J. M. & Simon, H. A. A theory of historical discovery: the construction of componential models. *Machine Learning 1:1* (1986) 107–136.

[‡]Note how no hypotheses use belief 3 as it did not contribute to the erroneous inference. Like all assumption-based systems, STAHLp is a dependency-directed reasoner because it only hypothesizes revisions to premises on which the erroneous inference ultimately depends.

[††]Both systems ran successfully on several examples described in (Zytkow & Simon, 1986), replicating the reasoning of early phlogiston theorists (2 premise reactions), Gay-Lussac and Thenard on potash and potassium (4 reactions), Black on alkalines (10 reactions), Berthollet on chlorine (5 reactions), and Lavoisier on oxygen (6 reactions). In addition to the above substances, models for iron, charcoal, water, lime and others were inferred.