DARPA-SN-17-57 Proposal

Abstract: This paper proposes a five (5) step process to develop a *formalized social sciences dataset* to automate in whole or part the verification of scientific theories, hypotheses, and/or studies within the social and behavioral sciences per DARPA-SN-17-57.

The five step process specifies constraints and methods to formalize, assess, and define theories, hypotheses, and/or studies that exist in written format by (1) *formalizing and defining theoretical terms in a mathematically precise and rigorous way* suitable to model theoretic treatments; (2) *corroborating the legitimacy of written documents* by checking researchers, institutions, journals, citations; (3) using natural language parsing to tease out specific markers that should or do modify our credence levels appropriately with respect to the works within which they are used; (4) *making explicit the underlying logic* to determine logical consistency; and (5) *identifying how well the theory, study, or hypothesis coheres with other high-credence theories, studies, or hypotheses*.

This approach aims to be comprehensive, but several potential limitations will be discussed. Various methodological assumptions or considerations are made available in the Appendix.

Adam InTae Gerard rev. 1.0.0 – 6.12.18

DARPA-SN-17-57 Proposal

Adam InTae Gerard

1. Main Proposal

Ockham (Ockham.io) is a working name or title for this idea. Ockham is a proposed scientific theory, hypothesis, and study automated credence setter.

The scope of this project is to partly or fully automate the process of determining the validity, legitimacy, of scientific theories, hypotheses, and studies.

To be clear, the aim of this project does not include automated theorem proving (of say the deductive sentences of a mathematical theory) nor does it include novel theory generation nor even novel prediction generation (automated, novel, and accurate predictions made within a theory).

2. Working Details

This section will briefly overview each of the five stages that when combined allow for creating a dataset to partly or fully automate the process of determining the validity, legitimacy, of scientific theories, hypotheses, and studies using machine learning.

Typically, automated verification comprises assembling a significant amount of accurate data then training some system using that data via an algorithm or structured data set assembled from that data.

Given the abundance of machine learning techniques, data structure systems, and computational power it is the opinion of this author that the primary problem confounding the creation of just such an automated system consists in exactly (1) which data

to use and (2) how one shapes the data into a data set to be used to partly or fully automate the verification process.

2.1. Formalizing and Defining Theoretical Terms

This stage consists in *formalizing and defining theoretical terms in a mathematically precise and rigorous way* suitable to model theoretic treatments of scientific theories and theory classification per Suppes[1].

This stage accomplishes three primary aims: (1) to ensure that the terms defined are not semantically vague or ambiguous[2], operationalized to the extent that they are practically useful, and genuine; (2) specification of the underlying formal mathematical models to ensure the correctness of results[3]; and (3) to assist in accomplishing the other stages to follow.

Consider this randomly selected finding from a randomly selected psychology article:

(S) "Performance was better for the elaborative-interrogation group than for the control group (76% versus 69%), even after controlling for prior knowledge and verbal ability."[4]

The sentence above is a declarative sentence – e.g. a grammatically correct, meaningful, assertion expressing a state of affair. Such a sentence is exactly the kind of sentence that predicate logics were originally invented to capture. Translation schemes of that sort are very familiar – Carnap[5], Prolog, and the wider view called the Syntactic View of scientific theories[5] come to mind. Indeed, in *model theory*, a

[1] See da Costa 2008 pp. 5 and Suppes 2002.
[2] Imprecision in theoretical terms means that in two different instances we might actually be talking about two very different things despite using the same word. In such cases, the utility of the terms of used is highly dubious as are any results that might be derived from them.
[3] See Suppes 2002 pp. 30-34 for a brief description of this method.
[4] Randomly from Dunlosky, John, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathen, and Daniel T. Willingham 2013. [5] See Halvorson 2016.
[5] See Appendix and Halvorson 2016.

theory (such as a scientific theory or hypothesis) is defined as a set of grammatically correct sentences in some logic *L*. After defining our predicates and constants in a formal logic:

(1) $E =_{df}$ Constant | elaborative-interrogation group | $E \in \Omega$

(2) $C =_{df}$ Constant | control group | $C \in \Omega$

(3) $P =_{df}$ Function Predicate | Performance | $\Omega \rightarrow [0,1]$

We can express the declarative sentence S above as:

(*) $P(C) < P(E)$ or (**) $P(C) = .69 \ \& \ P(E) = .76$

While simple, it should suffice to demonstrate how such a translation scheme is usually employed. Here, we go one step further and follow Suppes in terms of applying our First-Order expression above to the construction of a set-theoretic structure suitable as a model. This gives rise to a compact object we can use throughout the remainder of the stages.[6]

2.2. Soft Verification

Corroborating the legitimacy of written documents by checking researchers, institutions, journals, citations.

Before detailing this stage further, it's worth it to review an important and related *fallacy* (an error of reasoning, logic, or cognition).

A way of presenting the *Fallacy of Appealing to Authority*:

(1) Person *P* is or is perceived to be an (epistemic) authority or expert about some topic *X*.

(2) *P* says that a hypothesis, study, or theory *T* within the topic *X* is true or correct.

[6] See Suppes 2002 and "The Set-Theoretic Conception Of Science" 2007 for an example of such an approach.

(C) Therefore, *T* is true or correct.

It's important to note that the connection between (1), (2), and the inference to (C) presented above is a *necessary* connection. Independent of evidence, the inference above asserts, just because someone is perceived to be an authority (which they might not be) and just because that person says something is true makes it true.

Please note that we are not committing that fallacy here. Instead, we will leverage external markers of credibility to assist in determining the *credibility* (or *probability* of truth) of a theory, hypothesis, or study.

In general, reputable, or prestigious institutions create positive feedback loops to regulate and check the reputability and prestige of the studies commissioned by them. The reputation and prestige of that institution largely derives from the quality, novelty, and accuracy of the work that institution has produced.

2.3 Natural Language Processing

Using natural language parsing to tease out specific markers that should or do modify our credence levels appropriately with respect to the works within which they are used.

This accomplishes or can help to accomplish two major objectives: (1) to in part build a scientific lexicon or encyclopedia of terms (each of which in turn can be scored according to credibility) and (2) to help establish the credibility of the theory, hypothesis, or study Π by determining the usage of those terms within the written or spoken media items in which Π appears. We must be careful to separate mere reference or allusion to a non-credible term from the use of such a term as a *theoretical term*.

The following three examples illustrate these uses and differences:

(1) Aether

(2) "Aether was a silly concept…"

(3) "We propose Aether as an explanation…"

(1) is the term, (2) is a mere allusion, (3) is an assertion of a theory, hypothesis, or study – i.e. a *theoretical term*.

2.4. Logic and Consistency Checking

Making explicit the underlying logic to determine logical consistency.

A minimum requirement should be that whatever logic is embedded into whatever mathematical theory that is determined per **2.1** is logically consistent with respect to its own logical axioms, inference rules, and semantics.

Quantum Logic undergirds Quantum Mechanics, Bayesian Logic is the foundation for much of Probability Theory, etc.

Contradictions discovered in this process will dramatically decrease the credibility of the theory, hypothesis, and study.

2.5 Coherence With Other (High-Credence) Theories

Identifying how well the theory, study, or hypothesis coheres with other high-credence theories, studies, or hypotheses.

Given the construction of compact set-theoretic objects outlined in **2.1**, it is possible to see how well the assertions of one theory, hypothesis, or study coheres with the assertions of other theories, hypotheses, or studies. While coherence as a *sole foundation for justification* is dubious due to so-called impossibility results[7], coherence as an add-on or additional

[7] See Meijs and Douven 2007 for an introduction to the debate and replies.

level of scrutiny or as assessment filter is widely used to check or verify scientific discoveries[8].

In this way, tensions between theories, hypotheses, or studies can be identified and credence levels can be modified in accordance with the level of compatibility a theory, hypothesis, or study has with the (up-to) the rest and more interestingly with other high-credence theories, hypotheses, or studies.

3.0 Building the Dataset

A dataset combining these different stages can then be constructed with the intent to (in full or in part) automate credence level setting for theories, hypotheses, and studies:

Very early sketch of a simple relational database implementation with data structures, models, and relations:

STRUCT_ID | DOMAIN | RELTIONS | LOGIC | FK_LIST_T_ID
1 | AE | F={1,3},G=DxD | CLASSICAL | 1,3,4

S_ID | TYPE | STRUCT_ID | CREDENCE | RESOURCE | NAME | FK_R_ID | FK_I_ID
1 | STUDY | 1 | .8 | https://... | “Studies of Normal People…” | 3 | 5

T_ID | TERM | CREDENCE | PREDICATE_CONSTANT
1 | AETHER | 0 | AE

R_ID | NAME | FK_ID_ID | CREDENCE 1
| Joe Everyman | 1 | .5

I_ID | INSTITUTION | CREDENCE

[8] See Thagard 2007 pp 28. While the exact terms ‘coheres’ or ‘coherence’ are perhaps not explicitly said by many scientists, the concept that a theory, study, or hypothesis should be logically consistent with our other best current scientific theories is often mentioned though not in those exact words. Theories, hypotheses, or studies that are not logically consistent warrant additional scrutiny and reservation. Consider the recent claim that the speed light had been broken – see Condliffe 2012.

1 | Harvard | .9

Fine-tuning (correcting, modifying) the data will be done manually for each *resulting stage*. Each of those fine-tuning operations will be audited for record-keeping and to assist in the automation of future fine-tuning operations.

Human intervention into correcting credence is encouraged until any errors in credence setting are ruled-out and enough accurate data would exist that the process could be largely automated with intervention being made on an as needed basis.

4.0 Automated Credence Setting

The specific formula for calculating the credence of a specific theory, hypothesis, or study τ would take the form:

$$C(\tau) = \mathrm{COH}(\tau) \circ \mathrm{SV}(\tau) \circ \mathrm{T}(\tau) \circ \mathrm{LOGIC}(\tau)$$

Where each '$\circ$' specifies an operation, 'C' denotes the overall credence function for τ, 'COH' denotes the to be specified coherence operation, 'SV' names the soft verification operation modifying the resultant credence by checking the credence levels of participant researchers and institutions, 'T' denotes any modification as a result of non-credible term use, and 'LOGIC' modifies the overall credence by checking logical consistency (presumably removing the credibility of the theory entirely if it isn't logically consistent).

Theories with little credibility could have a negative impact on the credibility of the individual researchers and the institutions that sponsored the work.

5.0 Approaches and Algorithms

Specific details about the actual algorithm are yet somewhat imprecise and TBD (pending wider and deeper investigation). However, some progress has been made pertaining to potentially means of implementation.

There are two broad classes of algorithms that can be introduced to assist in (semi-)automated verification of scientific theories:

(1) ***Credibility Markets*** – so-called credibility markets regard the exchange credibility-backed or based goods (expertise, quality, credit-worthiness, asset rating, etc.) within a (capital) marketplace.[9]

In such markets, various policy mechanisms have been tested to determine impact and result on overall consumer purchasing and consumption, likelihood of sale, etc.

(2) ***Schelling (Focal) Points*** – a Schelling point refers to a coordination strategy in information-poor, partly secretive, or outcome-unknown circumstances (including jury voting, trust, street car traffic stops, credibility, etc.).[10] Such focal points often involving consideration of the mirroring or coordination strategy itself by agents within such games.

While somewhat high-level and abstract, both systems can be implemented and in several ways. Such implementations then solve to the correct algorithm (generating it dynamically) from some supplied data-set.

In the first case, each step of the multi-step approach laid out in the preceding sections can be tested against some control then compared against an independently scored data set to determine correctness of assigned credibility.

[9] See Frenkel, Jacob and Guillermo A. Cavlo 1991.
[10] See Janssen 2006.

In the second case, the coordination problem involves correctly rewarding credible empirical research and punishing less credible research. *Replication* (or the ability for other scientists and researches to recreate the results of an experiment) could be the *focal point itself*.

Both approaches can be undertaken *simultaneously* to increase the probability of success.

6.0 Reduced Limitations, Continuing Considerations

Many specifics regarding the implementation of an algorithm (or class of algorithms) remain opaque and TBD though the nature of such approaches is partly improved.

Refusal or deliberate unwillingness to format papers or studies in a manner amenable to this project could be made sanctionable.

It is the hope of this author that the dataset and overarching system above would be open-sourced and made publicly available for scrutiny, improvement, and to aid in public policy decision-making. Blockchain technologies represent a potential way to bring Ockham to life while keeping making all data public through the blockchain ledger, distributing the data for redundancy, and leveraging distributed computing to perform the actual calculations (which are relatively minor).

Such a blockchain implementation might provide a self-sufficient source of funding, economic incentives for participation, etc.

Auditing every *n* changes can help to get a clear understanding of what's going on under the hood. Neural Networks are considered black boxes within which very little can be discerned. That's no entirely correct. Each simulated activation function and neuron output can be recorded and logged (though this would be tedious and potentially storage massive).

7.0 Works Cited

Alston, William P. "Sellars and the 'Myth of the Given'." Philosophy and Phenomenological Research, 65 no. 1, July 2002. ⟨https://www.jstor.org/stable/3071107⟩.

da Costa, Newton C.A. "The mathematical role of time and space-time in classical physics." *Arxiv*, 7 February 2008. ⟨https://arxiv.org/pdf/gr-qc/0102107.pdf⟩.

Dunlosky, John, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathen, and Daniel T. Willingham. "Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology." *Association for Psychological Science*, vol. 14, no. 1, 2013, pp. 4-58. ⟨http://www.indiana.edu/~pcl/rgoldsto/courses/dunloskyimprovinglearning.pdf⟩.

Condliffe, Jamie. "Did Scientists Really Just Break the Speed of Light?" Gizmodo. 7 May, 2012. ⟨https://gizmodo.com/5908206/did-scientists-really-just-break-the-speed-of-light⟩.

Frenkel, Jacob and Guillermo A. Cavlo. "Credit Markets, Credibility, and Economic Transformation." *Journal of Economic Perspectives* vol. 5, no. 4, 1991, pp. 139-144. ⟨https://pubs.aeaweb.org/doi/pdf/10.1257/jep.5.4.139⟩.

Halvorson, Hans. "Scientific Theories." *The Oxford Handbook of Philosophy of Science*. Paul Humphreys ed. Oxford: Oxford University Press, 2016. ⟨http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199368815.001.0001/oxfordhb-9780199368815-e-33⟩. ⟨http://philsci-archive.pitt.edu/11347/1/sci-theories.pdf⟩. (preprint)

Janssen, Maarten C. "On The Strategic Use Of Focal Points In Bargaining Situations." Journal of Economic Psychology, vol. 27, no. 5, October 2006, pp. 622-634.

⟨https://www.sciencedirect.com/science/article/abs/pii/S0167487006000481⟩.

⟨https://papers.tinbergen.nl/06040.pdf⟩.

Meijs, Wouter and Igor Douven. "On the Alleged Impossibility of Coherence." *Synthese*, vol. 157, no. 3, 2007, pp. 347-360.

⟨https://ai2-s2pdfs.s3.amazonaws.com/8e51/dcf5747d7e1e20df03823e73303f9c2806b3.pdf⟩.

Meyerson, Roger B. "Learning From Schelling's 'Strategy of Conflict'."

⟨http://home.uchicago.edu/rmyerson/research/stratofc.pdf⟩.

Suppes, Patrick. *Representation and Invariance of Scientific Structures*. Stanford, CA: CSLI, 2002.

⟨http://web.stanford.edu/group/cslipublications/cslipublications/pdf/1575863332.rissbook.pdf⟩.

Thagard, Paul. "Coherence, Truth, and Development of Scientific Knowledge." *Philosophy of Science*, 2007, vol. 74, pp. 28-47.

⟨http://cogsci.uwaterloo.ca/Articles/coherence.truth.pos.2007.pdf⟩.

"The Set-Theoretic Conception Of Science." *Scientific Progress. Synthese Library* vol. 153, 2007, pp. 108-122.

⟨https://link.springer.com/chapter/10.1007/978-1-4020-6354-1_11⟩.

8.0 Appendix

I. Key Terms

Epistemic Credence: the level of *trustworthiness* or probability that we think a proposition, hypothesis,

declarative sentence, or theory ∏ is true. Tied to the level of justification we think that ∏ has.

Model Theory: the study of set-theoretic semantic structures that make all the sentences of formal language true.

Dataset: a set of mappings *D* between an input(s) and an output(s) such that given enough accurate information and time, reasonably accurate and consistent predications can be made by a neural network when trained on *D*.

Neural Network: an implementation of a biological neural network such that artificial neurons are connected into layers that are in turn connected to each other. Inputs are passed in to the first layer of neurons, functions applied to those inputs such that an output is produced by the last layer of the network.

Formalization: here, to *translate* or *transpile* a *fragment* of a *natural language* (up to and including the natural language itself) into a *formal language*.

Formal Language: A mathematically rigorous language used to, without ambiguity or vagueness, specify a domain (topic and scope of inquiry), artificial language, rules of inference, and truth-conditions. A *logic*, a *grammar*, and a *semantics*.

Nonmonotonic Inference: a type of non-classical inference whose consequent or conclusion is altered upon addition or deletion of information. *Abductive reasoning* a paragon example of this type of reasoning.

II. Sciences of the Psyche

Psychiatry: the medical practice of diagnosing, identifying, and treating mental illness by distinguishing underlying physiological, neurological, chemical, anatomical, or biological conditions, phenomena, or causes.

Psychology: the non-medical study of mental phenomena research and practice of which is usually limited to correlational studies and clinical counseling.

Neuroscience: the scientific study of the brain and its operation by recourse to biological, anatomical, computational, linguistic, and chemical explanations.

Cognitive Science: the application of computational, formal, mathematical, and linguistic methods to the study of mind, thought, and brain.

III. Key Relevant Philosophy Concepts

Semantic View of Scientific Theories: a scientific theory should be conceived as the *model class* of a set of grammatically valid sentences of some formal language.

Syntactic View of Scientific Theories: a scientific theory should be conceived as a set of grammatically valid sentences of some formal language.

Epistemic Scientific Structuralism: the position that our knowledge of the world is limited to structural knowledge of the world (while remaining silent about or denying the possibility of knowledge of *things-in-of-themselves*).

Connectionist Functionalism: *functionalism* is not to be confused with *symbolic computationalism*, *functionalism* is merely the thesis that (1) mental states are either identified by what they do rather than what they are made of (e.g. - some substance, *haeccity*, or primitive *thisness*) and/or (2) determined by the role they play or the system of which they are part (and hence, could be implemented in numerous substrates). The main version of this view is *connectionist functionalism* which identifies the functional systems giving to rise to mental phenomena with neural networks.

A priori: what can be known without experience – unknown theorems of mathematics are just such examples.

A posteriori: that which cannot be known *a priori* – e.g. Empirical science requiring observation of some physical phenomena – discovering a new species, a new exoplanet, or force of nature.

IV. Relevant Considerations

The utility of correlations: Discuss real versus fake patterns (or recast as good patterns).

Logical positivism and the problem of reducing scientific statements to sense data or observables:

There have been numerous issues pertaining to identifying or reducing high-level scientific terms to underlying primitive sense-datum or other observables. Some of these philosophical worries have bled into other fields under moniker "Symbol Grounding" (per the "Symbol Grounding" problem in Artificial Intelligence).

I do not have room here to discuss this too deeply but I will briefly say that there are at least three problematic assumptions interwoven into such concerns that give rise to the problem in the fist place: (1) Noumena / Phenomena distinction and the picture of representation that arises from it[11], (2) reductivism grounded in object-based ontologies[12],

[11] Some of these worries fall away when metaphysics is recast along modern mathematical lines. Isomorphism (or even a weaker morphism), Univalent Foundations (which equates *identity* and *isomorphism* as a fundamental axiom), and non-objectual ontologies help to blur these two Kantian distinctions. There is not so much an inner and outer (mental and external world), per say, but rather two layers of roughly approximate structural data (forthcoming).

[12] Which necessitates that there are some "Given" empirical primitives (see Alston 2002) or atoms that all sensory experience is ultimately or rests upon foundationally.

and (3) the Cartesian dogma of inner 1st person mental experience and outer 3rd person reality.

This proposal is not reliant on dogmatic foundational terms and such considerations are at this point largely irrelevant.

Limitations of Artificial Intelligence and Machine Learning: it takes time to develop a robust dataset that can be trained on a neural network to reliably and accurately generate correct outputs.

V. Professional Qualifications

I have a humble background in philosophy and software.

For a full summary of my professional and academic history please look at my LinkedIn.

A resume has also been attached.