Issues in Information Systems
Volume 23, Issue 3, pp. 100-113, 2022

DOI: https://doi.org/10.48009/3_iis_2022_109

A machine-learning approach for research model creation in theory building research

Ling Jiang, *York University, lnjiang@yorku.ca*
Christian Wagner, *City University of Hong Kong, c.wagner@cityu.edu.hk*

Abstract

We propose and illustrate an automated process for research model creation and hypothesis generation. To do so we argue that the early phases of theory building research are algorithmic and can be automated, given that a corpus of knowledge from past research has been appropriately codified. A pilot implementation is illustrated, together with an argument for the feasibility of knowledge extraction from past research. When fully implemented, the approach could significantly disrupt the research process, increasing process speed and enhancing creativity.

Keywords: theory building, research model, machine intelligence, logic programming

Introduction

One of the core objectives of academic research is to explain phenomena based on research models with theoretical grounding and empirical substantiation. For example, we may observe a contradictory phenomenon, such as that people who are a highly intelligent (e.g., professors) are not correspondingly wealthy. Hence, we might formulate the question "*why are people smart and yet they are not rich*", or more generally, "*does intelligence (not) lead to wealth*?" From a research perspective, we thus identify wealth as dependent variable, intelligence as independent variable, and then search for explanations that would associate the two with each other, formulating a theory based explanatory model that can be subsequently tested (Strenze 2007, Zagorsky 2007, Judge et al. 2009).

The process that leads to theory building and the definition of empirical research questions (hypotheses) is difficult, time consuming, and potentially error prone (Handfield and Melnyk 1998, Rivard 2020). The researcher has to develop a command of relevant theories, typically through a (more or less) extensive literature review that identifies relevant theories (Schryen et al. 2021). The researcher then has to select the most promising theory or theories as underpinning to formulate a research model that explains the phenomenon at hand (see steps 3-5 in the flow shown in Figure 1). In doing so, the researcher has to analyze a potentially large corpus of research articles to extract likely explanations in an unbiased and rational way. Dealing with potentially competing or conflicting explanations, the researcher manages to provide justifications for all hypothesized causal relationships in the research model. Especially for early career researchers, the "literature review" is a dreaded activity, which appears to be never ending, while raising complexity as it proceeds (Larsen et al. 2019, Cram et al. 2020). Seasoned researchers who have a repertoire of well-known theories may be prone to forego further search and become victims of selection biases, i.e., choosing well-established explanations for new phenomena. Consequently, an automated approach to research model creation and hypothesis generation that has access to a large number of research models can

significantly alter the theory building process (Lindberg 2020, Johnson et al. 2021, Wagner et al. 2021). An automated approach would choose based on best fit (not researcher preference), would suggest rival research models or model variations, and could carry out an exhaustive search of the research corpus in relatively short real time, not weeks or months. Hence it could significantly disrupt traditional research processes by vastly increasing speed, research model diversity, and model robustness.

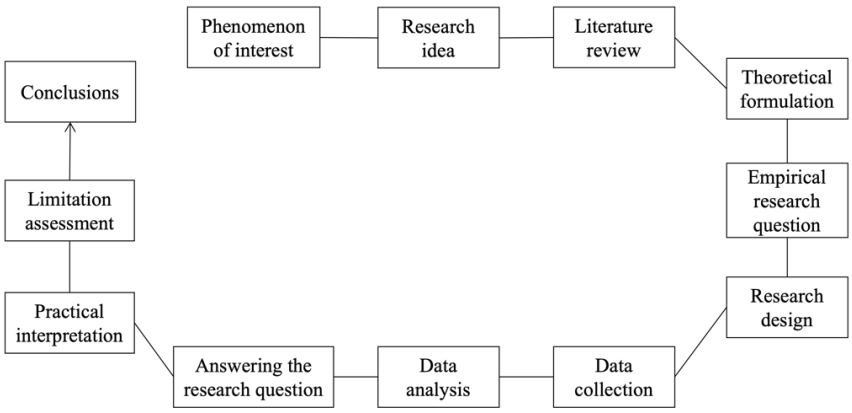

Figure 1: Theory Building Research Process

Given this challenge and opportunity, the question arises whether the literature review, theoretical formulation, and empirical research question activities can be turned into an automated or semi-automated process that would save time, reduce complexity, and avoid or reduce bias. Exploring this question and suggesting possible answers is the focus of this article.

In the article we will argue that the early steps of the theory building research process can be algorithmized and automated (Johnson et al. 2021, Wagner et al. 2021). To do so will hinge on a process that would capture the knowledge of a large corpus of research work in a well-defined (machine interpretable) form and thus make it amenable to collection, selection, combination, and questioning. The computational techniques of machine learning, which refers to tools, methods, and techniques for learning and improving task performance with experience (Goodfellow et al. 2016), enable the automation of theory building activities that are performed manually by researchers (Marshall and Wallace 2019). In contrast to existing machine learning approaches heavily relying on statistical models (Thilakaratne et al. 2020), this study proposes a knowledge-based approach to extracting and discovering knowledge from theory building research articles through logical reasoning. Building upon logic programming using Prolog language, we conceive a logical formalism to represent scientific knowledge in a structured manner (i.e., a knowledge base), and define logical rules to construct a reasoning engine capable of retrieving existing knowledge and deriving new conclusions from the knowledge base.

Using the above-mentioned example of exploring the relationship between intelligence and wealth (i.e., Judge et al. 2009), we build a prototype (a Prolog system) to illustrate how such a research corpus can be, and then be used through a reasoning engine to support research model and hypothesis formulation. We further demonstrate how multiple research models on the same topic can be synthesized into a united research corpus, upon which knowledge reasoning and inquiry can be carried out.

Background on the theory building research process

Theory can be broadly regarded as any coherent description or explanation of observed or experienced phenomena (Gioia and Pitre 1990). By definition, a theory is "an explained set of conceptual relationships" (Bunge 1967). The formal conceptual definition of theory consists of four components: (1) definitions of terms that used in the theory (Who? and What?), (2) a domain (e.g., setting or circumstance) where the theory can be applied (When? and Where?), (3) a set of relationships to explain how and why terms interact with each other (How? and Why?), (4) predictions for future happenings (Would? Should? and Could?) (Wacker 1998, 2008). Theory building refers to the process or cycle by which such an explained set of conceptual relationships are generated, tested, and refined. Theory building research is paradigmatically anchored according to philosophical views about the nature of ways of studying phenomena. To accommodate different philosophical views and conceptual paradigms, Burrell and Morgan (1979) have proposed a 2*2 matrix along two dimensions—objective-subjective and regulation-radical change, yielding four different research paradigms for theory building: radical humanist, radical structuralist, interpretivist, and functionalist. Among the four paradigms, our study focuses on the functionalist paradigm, which is characterized by an objectivist view with an orientation toward stability or maintenance of the status quo.

Theory building research grounded in functionalist paradigm aims to uncover universal regularities and relationships underlying observed or experienced phenomena. The process of theory building typically follows a deductive approach characterized by five key phases: phenomenon understanding, literature review, construct analysis, hypothesis formulation, and hypothesis validation (Handfield and Melnyk 1998, Holton III and Lowe 2007). At the outset scholars spot an area for research by exploring counterintuitive or abnormal phenomena. To fully understand the existing body of knowledge on the research area, researchers conduct a systematic review of literature related to the phenomenon of interest. Through construct analysis that summarizes the constructs used in the literature and their roles (i.e., antecedent, outcome, mediator, moderator), researchers determine the key constructs that should be included in theory building. Deducing from existing theories, a research model is created in a form of a set of hypotheses, coupled with justifications for the logical relationships among constructs. The validity of the research model is tested through empirical data measuring the constructs. The newly created research model that survives the empirical test claims its theoretical contributions by refining or extending existing theory.

In practice, the validated research model usually finds its expression in the article title, abstract, model, or validated hypotheses statements. For example, intrigued by the presumption of "the inside that accounts", Judge et al. (2009) started with an unsettled question of how "brains and beauty" account for success in life and wealth in particular. Prior research has identified key predictors of income, such as intelligence, physical attractiveness, and self-views. As the result of construct analysis, Judge et al. (2009) included six variables, i.e., general mental ability, physical attractiveness, education attainment core self-evaluations, income, financial stains, as the theory building blocks. To explain the mechanisms by which the inside (intelligence) and the outside (physical attractiveness) influence income and financial strains, a research model was formulated via a set of hypotheses (see Appendix A). The proposed hypotheses were empirically tested and supported by a longitudinal survey in the U.S.

Contributions to theory building research

This paper contributes to theory building research on three fronts. First, complementing to the statistical, probabilistic, and co-occurrence models (Marshall and Wallace 2019, Thilakaratne et al. 2020, van Dinter et al. 2021) for automating building theory research, this paper algorithmizes and automates the theory building research process by proposing a knowledge-based approach to knowledge extraction and representation. Regarding theory building research, our approach lays out a systematic scheme to extract knowledge components and designs a logical formalism to represent the knowledge components. Second,

by means of logic programing (Bratko 2000, Thaicharoen et al. 2009), we devise a reasoning engine that is able to allow machine to store, retrieve, reason, and synthesize the knowledge extracted from theory building research. Instead of constraining to a specific domain such as biomedical research (Kastrin and Hristovski 2021), management and business science (Johnson et al. 2021), the rules built in the reasoning engine are generalizable to theory building research across disciplines. Third, apart from knowledge representation and inquiry (Jeyaraj and Dwivedi 2020), the recursive algorithms we develop for the reasoning engine enable logical learning and knowledge discovery that is otherwise invisible for researchers. The pilot implementation using logic programing illustrates the principle working mechanism of our knowledge-based approach, and more importantly, demonstrates the practical feasibility of the approach. With a fully implemented knowledge base containing a substantial number of research models, our approach could dramatically speed up the process of research development, exquisitely preserve the diversity in research models, and intellectually stimulate the creativity of knowledge discovery, thereby disrupting the conventional process for research model and hypothesis generation.

Knowledge-based approach

Knowledge Extraction from Theory Building Research

The deductive approach to conducting theory building research gives rise to a twofold process by which scholars deal with knowledge: (1) retrieving existing knowledge from prior research through literature review, and (2) creating new knowledge to advance existing theories through proposing and validating hypothesis. The newly created knowledge is subsequently added to the literature in a recursive way to broaden the knowledge repository, which remains open for future disapproval. To facilitate the process of knowledge retrieval and knowledge creation process, our study employs a knowledge-based approach to extract and represent knowledge embedded in theory building research in a systematic manner. Given the varying formats used to articulate knowledge in theory building research, as shown in Table 1, we extract the key knowledge entities and corresponding components. Central to theory building research is a research model, usually represented by a figure which connects boxes (constructs) via arrows representing causal links (e.g., the research model in Appendix A). In line with the definition of theory as "an explained set of conceptual relationships", we decompose a research model in a recursive sense that it comprises an array of hypotheses, each of which elaborates a relationship among constructs and is supported by justification.

Table 1: Knowledge Components in Theory Building for a Research Article

Entity	Component	Format
Research topic	Keywords	Text: Phrases (constructs)
	Research context	Text: Description of where and when the research model is applicable
Theoretical foundation	Existing theory	An array of hypothesis
	Prior research	Reference: An array of articles
Research model	Hypothesis	Text: Statement that articulates the relationship among constructs
		Figure: Multiple rectangles connected by arrow(s)
	Construct	Text: Conceptual objects referred in hypothesis
		Figure: Rectangle in cause-effect model
	Relationship	Text: Logical order described in hypothesis
		Figure: Arrow in cause-effect model
	Justification	Text: Argument supporting hypothesis (theoretical reasoning)
Empirical testing	Research method	Text
	Empirical results	Text
		Figure: Cause-effect model with statistical significance of hypothesis

Design of the Knowledge Representation

To make scientific knowledge understandable and computable for intelligent machine (Hunter and Liu 2010), we propose a knowledge representation scheme tailored to theory building research articles. As one of the most important subareas of artificial intelligence (AI), knowledge representation aims to understand, design, and implement ways of representing information such that a machine can process this information to derive new conclusions, to converse with people in natural languages, to make decisions, and to solve problems that normally require human expertise (Russell and Norvig 2010). Our study focuses on the logic-based approach to scientific knowledge representation and reasoning (Baader 1999, Lloyd 2012).

In the context of theory building research, the core of knowledge representation is logic (Moore 1982), which plays an essential role in formulating hypothesis, for instance, employing declarative sentences as a basic for representing common sense. Declarative sentences regard knowledge as descriptions of notations, facts, and logic rules that can be denoted through a set of predicates. Compared with representing knowledge in segments of computer programs or in tables, knowledge expressed via declarative sentences is modular and minimally dependent on the order of declaratives, thus can be used for multiple purposes across many domains (Baral and Gelfond 1994).

Prolog ("**Pro**gramming in **Log**ic") has been widely used for prototyping and building knowledge bases (Bratko 2000) through declarative logic. Prolog provides the facility to build declarative knowledge bases, paired with an inference engine to evaluate queries against the knowledge bases to determine their truth. Knowledge represented by Prolog falls into two main types of declaratives: *fact* and *rule*. Whereas a fact is denoted as a base clause with an empty body (i.e., a primitive proposition without "if part"), a rule takes a form of a clause that has at most one positive literal (i.e., if-then proposition). A collection of Prolog facts and rules constitutes a knowledge base. Regarding a research model as an essential carrier of scientific knowledge, we represent a theory building research article as an array of facts, each of which corresponds to a hypothesis in the research model of the article.

According to the distinguishing nature of hypotheses, we propose a scheme to represent a research model through three types of facts (see Table 2): *direct_hypothesis*, *mediation*, and *moderation* declared in a Prolog program, to fully capture logical orders embedded in the research model. A predicate fact declaration decomposes hypothesis into multiple elements, each of which serves as an argument. Facts concerning direct hypotheses are defined by four arguments: dependent variable (consequent), independent variable (antecedent), type of relationship (either association or causality), and sign of relationship (either positive or negative). Arguments for mediation fact include dependent variable, independent variable, mediator, and the type of mediation (either full or partial). Moderation fact declares dependent variable, independent variable, moderator, and the effect of moderation (either strengthen or weaken). Table 2 also illustrates the three types of fact using Prolog examples. Formalizing hypotheses into Prolog facts allow us to establish knowledge base by synthesizing research models across theory building studies.

Table 2: Research Model Represented by a Prolog Program

Fact Type	Argument	Hypothesis Example	Prolog Program Example
Direct hypothesis	• Dependent variable • Independent variable • Type of relationship (association, causality) • Sign of relationship (positive, negative)	General mental ability is positively associated with income.	**direct_hypothesis**(income, general_mental_ability, association, positive).
Mediation	• Dependent variable • Independent variable • Mediator • Type of mediation (partial, full)	Educational attainment partially mediates the relationship between general mental ability and income.	**mediation**(income, general_mental_ability, educational_attainment, partial).
Moderation	• Dependent variable • Independent variable • Moderator • Effect of moderation (strengthen, weaken)	Age moderates the relationship between income and physical attractiveness such that the relationship is stronger at an earlier age*.	**moderation**(income, physical_attractiveness, age, weaken).*
*Note: The statement is proposed as an example of moderation, rather than a formal hypothesis investigated in Judge et al. (2009).			

Reasoning Mechanism

The capability of reasoning with a knowledge representation scheme manifests the intelligence of a machine. While declarative reading of logic program informs knowledge representation, procedural interpretation of logic program empowers automated reasoning engine (Kowalski 1974). Applying three typical reasoning approaches, i.e., deduction and induction (Minnameier 2010, Rodrigues 2011), we devise a set of inference rules to build a reasoning mechanism, which together with Prolog's general inference engine, provides the machine intelligence for our prototype.

Presupposing the existence of truth and falsity, deduction draws logical consequences from premises. The truth of an inference relies on the truth of all premises (Evans et al. 1993). A basic rule of inference in deduction is *Modus Ponens*, i.e., the fallacy of affirming the consequent (Bobzien 2002). For instance, given a premise "If rains, then the lawn is wet" and the fact that it rains, we can draw a conclusion that the lawn is wet. Induction, on the contrary, concludes a general law based on numerous empirical observations (Arthur 1994, Feeney and Heit 2007). The relative frequency of a specific pattern indicates its probability of being true. Once a pattern is identified through evidence in large scale, it is added into the existing body of knowledge and applied to new case. For example, whenever it rains, the lawn is wet. When the coincident has been observed for 1000 times, the law of "If rains, then the lawn is wet" is concluded, such that for the next time it rains the lawn must be wet. More generally speaking, the assumption is that a phenomenon having been observed consistently n times, it should also be observed in the $n+1^{th}$ incidence. The belief strength should obviously rise with increasing size of n.

Scientific research seeks to unveil truth beneath phenomenon observed. Constrained by the economy of research, however, it is infeasible to falsify all possible explanations to let the most viable one stand out as an approximation to truth. Concerned with such a research dilemma, the purpose of our reasoning engine is twofold: 1) to identify inconsistent propositions in the literature, and 2) to tease out promising truth through logical reasoning with knowledge base. Our knowledge representation scheme lays the foundation for building such a reasoning engine to automatically derive conclusions implied by the knowledge already

present in knowledge base, herein the facts representing research models. The declarative logic programing for knowledge representation enables us to detect inconsistent facts and to discover implicit knowledge through propositional resolution, a rule of inference for propositional logic that can be used for both proving inconsistence (i.e., refutation resolution) and deriving new conclusion (i.e., direct resolution) (Stickel 1985).

The knowledge base may represent research models from multiple studies, making it possible that hypotheses from different research sources contradict with each other logically. Refutation resolution is sound and complete to identify unsatisfiable consequences derived from the set of facts (Kifer and Lozinskii 1992). Besides the inconsistency residing in predicate formalism, relationships proposed by different articles might not be consistent due to controversy and debate in academic conversation. For instance, a typical inconsistency we encounter during literature review is that both positive and negative impacts (i.e., contradictory signs) are hypothesized between two constructs. Based on the set of logic clauses declared in the knowledge base, we develop "if-then" rules (see Table 3) to instruct reasoning engine to detect contradictory relationships. By virtue of thorough detection of inconsistency underlying the knowledge base, our reasoning engine is able to chart the literature landscape for a given knowledge domain in a logical and systematic manner.

Apart from discerning logic inconsistency through refutation resolution, the reasoning engine employs the logic of deduction and induction to derive implicit knowledge underneath the facts through direct resolution (see "if-then" rules in Table 3). Specifically, the logical representation of knowledge facilitates the deduction of inferences through automating resolution among first-order clauses (i.e., facts). In light of the declaration of arguments for each type of fact in Table 2, reasoning engine collects all constructs, as well as their roles in terms of antecedent, consequent, mediator, moderator, upfront antecedent without any preceding antecedent, and backend consequent without any subsequent consequent. Transitive relation signals implicit mediation among interconnected direct hypotheses. More importantly, logic programing supports recursion, thereby enabling the reasoning engine to exhaust implicit mediations connecting a chain of mediators (i.e., the degree of separation). Meanwhile, the collection of hypotheses in the knowledge base prepares the research corpus for inducing intensively studied constructs and relationships, which are indicated by the frequency. Given the set of consequents declared in the knowledge base, all antecedents to the concerned consequent serve as plausible explanations. Among the resultant plausible antecedents, the frequency of each candidate implies the credibility of being the best explanation within the knowledge base.

Table 3: Rules for the Reasoning Engine

Approach	Purpose	Rule
Refutation	Detect inconsistency in terms of relationship sign, causal interpretation, and moderation effect	***If*** A→B and different signs exist, ***then*** inconsistency
		If A→B and B→A and signs are opposite, ***then*** inconsistency
		If A→B and B→A and both are causality, ***then*** inconsistency
		If A→B and B→A and both association and causality exist, ***then*** inconsistency
		If A→B moderated by C and different effects exist, ***then*** inconsistency
Deduction	Collect construct and identify its nature (e.g., antecedent, consequent, mediator, and moderator)	***If*** A→B or C mediates A→B or C moderates A→B, ***then*** A, B and C are constructs
		If A→B, ***then*** A is an antecedent
		If A→B, ***then*** B is a consequent
		If A→B and C mediates A→B, ***then*** C is a mediator
		If A is an antecedent and consequent but not a mediator, ***then*** C is an implicit mediator
		If A→B and C moderates A→B, ***then*** C is a moderator

Table 3 (Continued)

Approach	Purpose	Rule
Deduction	Identify upfront antecedent and backend consequent	***If*** A is an antecedent and A is not a consequent, ***then*** A is an upfront antecedent
		If A is a consequent and A is not an antecedent, ***then*** A is a backend consequent
	Explore potential mediation through transitive relations among direct hypotheses	***If*** A→B and B→C, ***then*** B mediates A→C with degree of separation at 1
		If $n > 1$ and B mediates A→C degree of separation at $n-1$ and C→D, ***then*** B and C mediate A→D with degree of separation at n
Induction	Identify intensively studied antecedent, consequent, mediator, and moderator based on construct frequency	***If*** A is declared as a construct and appears in relationships as independent variable with high frequency, ***then*** A is an intensively studied antecedent
		If A is declared as a construct and appears in relationships as dependent variable with high frequency, ***then*** A is an intensively studied consequent
		If A is declared as a construct and appears in mediation as mediator with high frequency, ***then*** A is an intensively studied mediator
		If A is declared as a construct and appears in moderation as moderator with high frequency, ***then*** A is an intensively studied moderator
	Identify intensively studied relationship based on relationship frequency	***If*** A→B as association appears with high frequency, ***then*** A is the core association
		If A→B as causality appears with high frequency, ***then*** A is the core causality
		If C mediates A→B appears with high frequency, ***then*** C mediates A→B is the core mediation
		If C moderates A→B appears with high frequency, ***then*** C moderates A→B is the core moderation

Notes:
- Expressions such as A→B refer to direct hypothesis proposed by research models.
- The arrow in expressions such as A→B accords with the arrow used in figure of research model, which is not necessarily an indicator of causal relationship.

Pilot implementation

Our reasoning engine equips knowledge base with rules (i.e., "if-then" rules in Table 3) to render a machine with intelligence to derive insights from the scientific literature. To enable the machine to interact with human, we demonstrate how to inquire knowledge by either retrieving clearly expressed knowledge or exploring undetermined questions. The duality of logic programming—as a formalism for knowledge representation and a computational mechanism for theorem proving, empowers intelligence machine to respond to human inquiry through computation on propositional logic (Avigad 2002). Based on the resolution principle for mechanical theorem proving, our prototype deals with knowledge inquiry through matching all facts and rules till the resolution reaches negation, which is approximate to unification in logic.

Explicit Knowledge Representation

Inquiries about existing knowledge search facts declared explicitly in the knowledge base, i.e., conceptual relationships hypothesized to be true by research models. Making such an inquiry means that the inference

engine (here: the Prolog interpreter) is asked whether facts can be found in the knowledge base that would make the inquiry true. If the answer is yes, the specific instantiation(s) that satisfy the inquiry will be returned. Otherwise, the inquiry fails. For instance, suppose a knowledge base includes the following four exemplar facts.

direct_hypothesis(income, general_mental_ability, association, positive).
direct_hypothesis(income, educational_attainment, association, positive).
mediation(income, general_mental_ability, educational_attainment, partial).
moderation(income, physical_attractiveness, age, weaken).

If we now wanted to inquire whether there was a relationship that associates income with general mental ability, we could simply issue a query using the (Prolog) inference engine. We would ask *?-* ***direct_hypothesis****(Y, X, T, S)*, with X, Y, T, and S representing variables which would make the inquiry true. With the knowledge base containing two direct hypotheses of this nature, two answers would be delivered by the system.

| ?- **direct_hypothesis**(Y, X, T, S).
Solution 1: Y = income, X = general_mental_ability, T = association, S = positive;
Solution 2: Y = income, X = educational_attainment, T = association, S = positive.

We note that the knowledge base also contains a fact to suggest that income and general mental ability are indirectly associated, mediated by educational attainment. To use this knowledge in a broader knowledge discovery process will require the definition of reasoning rules.

Knowledge Discovery

When it comes to discovering implicit knowledge, inquiries primarily take place on rules (see the "if-then" rules in Table 3) that govern the reasoning engine to derive knowledge implied by the knowledge base. According to Prolog syntax, "if-then" rule is defined as a full horn clause, namely *predicate*, which consists of two components: head ("then" part) and body ("if" part). The head of a clause declares a predicate using a unique name and argument specification. When one rule accommodates multiple logic scenarios, its predicate may consist of multiple clauses with the same head (name and argument list) but different bodies (logic). As the interface for inquiring implicit knowledge, predicates are a set of declarative sentences that translate logical rules into program via logic programming language. Table 4 illustrates how "if-then" rules are declared as Prolog program using exemplar rules in Table 3. As the predicate of "*mediator_implicit(M)*" defined in Table 4, for instance, we explore an implicit mediator by searching a construct fulfilling three criteria: 1) serving as an antecedent in one direct hypothesis, and 2) serving as a consequent in another direct hypothesis, 3) not being explicitly hypothesized as a mediator in any mediation relationship.

Table 4: Illustration of Prolog Predicate Implementing Reasoning Rules

Logical Rule	Predicate

If A→B or C mediates A→B or C moderates A→B, ***then*** A, B and C are constructs	**construct**(X) :-
	direct_hypothesis(X, _, _, _);
	direct_hypothesis(_, X, _, _);
	mediation(_, _, X, _);
	moderation(_, _, X, _).
	construct_distinct(X) :-
	distinct(**construct** (X)).

If A→B and B→C, ***then*** B mediates A→C with degree of separation at 1	**mediation_implicit**(Y, X, 1, association) :- direct_hypothesis(Y, Z, association, _), direct_hypothesis(Z, X, association, _), not(**mediation**(Y, X, Z, _)). **mediation_implicit**(Y, X, 1, causality) :- direct_hypothesis(Y, Z, causality, _), direct_hypothesis(Z, X, causality, _), not(**mediation**(Y, X, Z, _)).
If *n* > 1 and B mediates A→C degree of separation at *n-1* and C→D, ***then*** B and C mediate A→D with degree of separation at *n*	**mediation_implicit**(Y, X, Degree, association) :- Degree > 1, direct_hypothesis (Y, Z, association, S), DegreeMinus is Degree - 1, **mediation_implicit**(Z, X, DegreeMinus, association). **mediation_implicit**(Y, X, Degree, causality) :- Degree > 1, direct_hypothesis (Y, Z, causality, S), DegreeMinus is Degree - 1, **mediation_implicit**(Z, X, DegreeMinus, causality).

The syntax of inquiry for implicit knowledge is akin to that of inquiry for explicit fact by calling the predicates of rules. Consider the rules declared using Prolog in Table 4 and the four-fact knowledge base as example. We can list all distinct constructs in the knowledge base through the inquiry "**construct_distinct**(X)", receiving the following results.

| ?- **construct_distinct**(X).
Solution 1: X = income;
Solution 2: X = general_mental_ability;
Solution 3: X = educational_attainment;
Solution 4: X = age.

At the core of knowledge inquiry are reasoning rules predefined in the reasoning engine for knowledge discovery. Instructed by the set of rules, the reasoning engine conducts an exhaustive search to resolve inquiries about the knowledge base. Reasoning with different knowledge bases, our Prolog system is able to retrieve and discover knowledge in different domains.

Discussion

At this point we have demonstrated that a knowledge base of facts and rules describing a research model can be used to query that model, extract constructs, and create hypotheses. The knowledge base in our pilot was small, however, and drawn only from a single article. Hence it begs the question whether in general, well-articulated knowledge models can be extracted from research articles. We believe that not all research

articles are amenable to this approach, but that there is a significant body of research defined well enough to allow for the development of a research knowledge corpus. Formally this can only be demonstrated through an experiment where a set of research articles is selected and then subjected to knowledge extraction. This is a process that goes beyond the scope of this article. For a simple proof of concept, we identified the most cited articles on Google Scholar (May 8, 2022) whose titles included a casual (smart, rich) and a more formal (intelligent, socioeconomic status) expression of the research question. The articles are Strenze (2007) and Zagorsky (2007). Strenze states (in the abstract) that "*intelligence is a powerful predictor of success but, on the whole, not an overwhelmingly better predictor than parental SES or grades. Moderator analyses showed that the relationship between intelligence and success is dependent on the age*". Zagorsky explains in the abstract that "*each point increase in IQ test scores raises income by between $234 and $616 per year after holding a variety of factors constant. Regression results suggest no statistically distinguishable relationship between IQ scores and wealth.*" Thus, it is worth noting that (1) articles are explicit in expressing model relationships and thus amenable to formal knowledge representation, (2) the need for a dictionary of synonyms, (3) the need to track the chains of logical reasoning, and (4) the need to create representations for equivocal or seemingly inconsistent knowledge. After all, according to Zagorsky (2007), there was a relationship between IQ and income, but not with wealth. Furthermore, the development of a corpus of knowledge, at least for now, requires human intelligence to identify the research knowledge, and to properly interpret and thus code research findings. Fortunately, the process would be one where the knowledge is coded once to be (re-)used many times.

Conclusions

The early phases of any research process, all the way to the formulation of research hypotheses, have traditionally been a time- and thought-intensive search process relying on human expertise and effort. At the same time, the process also seemingly routine, relied on exhaustive search, and knowledge selection, rather than discovery. Our pilot implementation of a reasoning system for the creation of research models and research questions illustrates that the process at least can be computer-supported, and at best be carried out dominantly through automated reasoning. The process of research design thus can be sped up and carried out in unbiased fashion. The availability of a corpus of research findings can also answer interesting questions such as "*what do we definitively know based on past research findings*?" In addition to speeding up modeling and hypothesis generation, a research corpus could also be used for knowledge discovery and exploration. An automated approach could thus change the research model building process from a more mundane search and selection process to a creative discovery process, thus enhancing not only the speed but also the originality of future research endeavors.

References

Arthur WB (1994) Inductive reasoning and bounded rationality. *American Economic Review* 84(2):406-411.

Avigad J. (2002) Logic and computation. https://www.andrew.cmu.edu/user/avigad/Teaching/landc_notes.pdf.

Baader F (1999) Logic-based knowledge representation. *Artificial Intelligence Today* LNAI 1600:13-41.

Baral C, Gelfond M (1994) Logic programming and knowledge representation. *Journal of Logic Programming* 19/20:73-148.

Bobzien S (2002) The development of modus ponens in antiquity: From aristotle to the 2nd century ad. *Phronesis* 47(4):359-394.

Bratko I (2000) *Prolog programming for artificial intelligence.* (Addison Wesley).

Bunge M (1967) *Scientific research: The search for system* (Springer-Verlag, New York).

Burrell G, Morgan G (1979) *Sociological paradigms and organizational analysis* (Heinemann, London).

Cram WA, Templier M, Pare G (2020) (re)considering the concept of literature review reproducibility. *Journal of the Association for Information Systems* 21(5):1103-1114.

Evans JSBT, Byrne RMJ, Newstead SE (1993) *Human reasoning: The psychology of deduction* (Lawrence Erlbaum Associates, Hillsdale, NJ).

Feeney A, Heit E (2007) *Inductive reasoning: Experimental, developmental, and computational approaches* (Cambridge University Press, New York).

Gioia DA, Pitre E (1990) Multiparadigm: Perspectives on theory building. *Academy of Management Review* 15(4):584-602.

Goodfellow I, Bengio Y, Courville A (2016) *Deep learning, adaptive computation and machine learning series* (MIT Press, Cambridge, MA).

Handfield RB, Melnyk SA (1998) The scientific theory-building process: A primer using the case of tqm. *Journal of Operations Management* 16(4):321-339.

Holton III EF, Lowe JS (2007) Toward a general research process for using dubin's theory building model. *Human Resource Development Review* 6(3):297-320.

Hunter A, Liu W (2010) A survey of formalisms for representing and reasoning with scientific knowledge. *Knowledge Engineering Review* 25(2):199-222.

Jeyaraj A, Dwivedi YK (2020) Meta-analysis in information systems research: Review and recommendations. *International Journal of Information Management* 55:102226.

Johnson CD, Bauer BC, Niederman F (2021) The automation of management and business science. *Academy of Management Perspectives* 35(2):292-309.

Judge TA, Hurst C, Simon LS (2009) Does it pay to be smart, attractive, or confident (or all three)? Relationships among general mental ability, physical attractiveness, core self-evaluations, and income. *Journal of Applied Psychology* 94(3):742-55.

Kastrin A, Hristovski D (2021) Scientometric analysis and knowledge mapping of literature-based discovery (1986–2020). *Scientometrics* 126(2):1415-1451.

Kifer M, Lozinskii EL (1992) A logic for reasoning with inconsistency. *Journal of Automated Reasoning* 9(2):179-215.

Kowalski RA (1974) Predicate logic as programming language. *IFIP Congress* (North-Holland), 569-574.

Larsen KR, Hovorka DS, Dennis AR, West JD (2019) Understanding the elephant: The discourse approach to boundary identification and corpus construction for theory review articles. *Journal of the Association for Information Systems* 20(7):887-927.

Lindberg A (2020) Developing theory through integrating human and machine pattern recognition. *Journal of the Association for Information Systems* 20(1):90-116.

Lloyd JW (2012) *Foundations of logic programming* (Springer, Heidelberg).

Marshall IJ, Wallace BC (2019) Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews* 8(1):163.

Minnameier G (2010) The logicality of abduction, deduction and induction. Mats Bergman, Sami Paavola, Ahti-Veikko Pietarinen, Rydenfelt H, eds. *Ideas in action: Proceedings of the applying peirce conference, nordic studies in pragmatism 1* (Nordic Pragmatism Network, Helsinki), 239-251.

Moore RC (1982) The role of logic in knowledge representation and commonsense reasoning. *The Second National Conference on Artificial Intelligence* (AAAI Press), 428-433.

Rivard S (2020) Theory building is neither an art nor a science. It is a craft. *Journal of Information Technology* 36(3):316-328.

Rodrigues CT (2011) The method of scientific discovery in peirce's philosophy: Deduction, induction, and abduction. *Logica Universalis* 5(1):127-164.

Russell SJ, Norvig P (2010) *Artificial intelligence. A modern approach* 3rd ed. (Pearson Education, New Jersey).

Schryen G, Wagner G, Benlian A, Paré G (2021) A knowledge development perspective on literature reviews: Validation of a new typology in the is field. *Communications of the Association for Information Systems* 49(1):134-186.

Stickel ME (1985) Automated deduction by theory resolution. *Journal of Automated Reasoning* 1:333-355.

Strenze T (2007) Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence* 35(5):401-426.

Thaicharoen S, Altman T, Gardiner K, Cios KJ (2009) Discovering relational knowledge from two disjoint sets of literatures using inductive logic programming. *IEEE Symposium on Computational Intelligence and Data Mining (CIDM'09)* (IEEE), 283-290.

Thilakaratne M, Falkner K, Atapattu T (2020) A systematic review on literature-based discovery. *ACM Computing Surveys* 52(6):1-34.

van Dinter R, Tekinerdogan B, Catal C (2021) Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology* 136:106589.

Wacker JG (1998) A definition of theory research guidelines for different theory-building research methods in operations management. *Journal of Operations Management* 16(4):361-385.

--- (2008) A conceptual understanding of requirements for theory-building research. *Journal of Supply Chain Management* 44(3):5-15.

Wagner G, Lukyanenko R, Paré G (2021) Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology* 37(2):209-226.

Zagorsky JL (2007) Do you have to be smart to be rich? The impact of iq on wealth, income and financial distress. *Intelligence* 35(5):489-501.

Appendix A: Research Model in Judge et al. (2009)