AN ANALYSIS OF INTERTHEORETICAL CONNECTIONS IN THE INTERDISCIPLINARY FIELD: SOME CASES OF COGNITIVE SCIENCE

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
der Ludwig-Maximilians-Universität
München

vorgelegt von

Steve Hendra

aus

Surabaya, Indonesien

2020

Referent: Prof. Dr. em. Dr. h.c. mult. Carlos Ulises Moulines

Korreferent: Prof. Dr. Bernhard Lauth

Tag der mündlichen Prüfung: 14.02.2020

Abstract

Background: Interdisciplinarity is one of the current trends in the scientific world today, that began with the uneasiness about the loss of the unity of science. This trend also opens possibilities for explaining complex phenomena more comprehensively and creating more advanced applications and implementations of scientific theories. One of the biggest challenges to conducting interdisciplinary research is theoretical integration, how can we combine theories from various disciplines such that the combination is fruitful?

Method: This dissertation attempts to answer this challenge by analyzing the intertheoretical connections of some theories from various disciplines for some real interdisciplinary research. The structuralist metatheory of science is applied as the basic theory to model the intertheoretical connections formally. This research begins with modeling the scientific theories in question before modeling the intertheoretical connections and some modifications needed. This research focuses on some researches in cognitive science that involve psychology, neuroscience, and artificial intelligence. The first research is the research conducted by van Veen et al., who research the activity of neurons in a brain's field called *dorsal Anterior Cingulate Cortex* during a phase of dissonance between cognitions. This research serves as a context for the analysis of intertheoretical reduction between the Festinger theory of cognitive dissonance and the Hawkins-Kandel computational neuroscientific theory. The other research is the consonance model of simulation built by Shultz and Lepper, which implements the Hopfield network to build a simulation of the cognitive dissonance. The third research is the connectionist model of simulation built by van Overwalle and Jordens, which implements the two-layers feed-forward perceptron and the delta rule as its learning rule to build a simulation of forced compliance dissonance.

Result: Through this research, the author concludes that the structuralist metatheory of science can be applied for modeling and analyzing intertheoretical connections in the same discipline and between disciplines in real scientific research. The structuralist metatheory of science enables us to model and analyze the structure of the theories and their intertheoretical connections with great detail and brings very fruitful results. This research delivers some results not only for the structuralist theory

science itself but also for the philosophy of science in general and interdisciplinary researches, especially cognitive science. First, by analyzing the models, a revision of the definition of intertheoretical specialization, and specialization of the concept of theory-holon according to the structuralist theory of science, called the V-pattern and strategy for combining scientific theories, are proposed. These V-pattern and strategy can serve as a tool for combining scientific theories.

Second, unlike other approaches on intertheoretical reduction such as the GNS model, the structuralist metatheory of science does not intend to formulate a generalized model of reduction nor focus only on intertheoretical reduction. It provides us powerful tools for modeling various intertheoretical relations, including intertheoretical reduction, case by case. Although the intertheoretical reduction in the structuralist model is more epistemological than ontological, the structuralist models show how the reduction has empirical claims and intended applications by applying the r* function that maps the T-theoretical level to T-non-theoretical level.

Third, related to the notion of the unity of science, this dissertation still sees that the unity of science is still a plausible and essential agenda for the philosophy of science and the scientific world in general. This dissertation's idea of the unity of science proposed does not assume essentialism, reductionism, and epistemological monism. This dissertation sees that the unity of science is closely related to scientific practice.

Fourth, for interdisciplinary research, primarily cognitive science, this dissertation proposes an approach to model and analyze intertheoretical connections for any scientific research or any philosophical school in philosophy of science related to the idea of intertheoretical relation. This dissertation is the first example of such modeling and analysis.

Eine Analyse der intertheoretischen Relationen in interdisziplinären Bereichen: Einige Fälle der Kognitionswissenschaft

Eine Zusammenfassung

"Interdisziplinarität" ist eines der wichtigsten Wörter, um einen Trend in der heutigen wissenschaftlichen Welt zu beschreiben. Dieser Trend der Interdisziplinarität begann mit Unbehagen über den Verlust der wissenschaftlichen Einheit am Anfang des 20. Jahrhundert und hat heute zugenommen. Dieser Trend eröffnet auch Möglichkeiten, komplexe Phänomene umfassender zu erklären und fortgeschrittenere Anwendungen und Implementierungen wissenschaftlicher Theorien zu schaffen. Laut Jungert spielen fünf Aspekte einer Disziplin eine wichtige Rolle bei der Untersuchung der Merkmale interdisziplinärer Beziehungen: Objekte, Methoden, Probleme, theoretischer Integrationsgrad und Personen/Institutionen (Jungert, 2013, S. 7–9).

Eine der größten Herausforderungen bei der Durchführung interdisziplinärer Forschungen sind die Fragen der theoretischen Integration. Wie können wir Theorien aus verschiedenen Disziplinen so kombinieren, dass die Kombination fruchtbar ist? Diese Dissertation versucht, diese Herausforderung durch Analyse der intertheoretischen Zusammenhänge einiger Theorien aus verschiedenen Disziplinen zu beantworten, die für einige echte interdisziplinäre Forschungen kombiniert werden. Diese Forschung konzentriert sich auf einige Forschungen in der Kognitionswissenschaft, die Psychologie, Neurowissenschaften und künstliche Intelligenz umfassen.

Um dieses Ziel zu erreichen, wird die strukturalistische Metatheorie der Wissenschaft angewendet, um die intertheoretischen Verbindungen

zwischen wissenschaftlichen Theorien aus verschiedenen Disziplinen durch Implementierung der Modelltheorie formal zu modellieren und zu analysieren. Ein Modell einer wissenschaftlichen Theorie wird als „Theorieelement“ bezeichnet und besteht aus einem Konzept des theoretischen Kerns und einer Menge von intendierten Anwendungen. Die strukturalistische Metatheorie der Wissenschaft modelliert die intertheoretischen Verbindungen als Beziehungen zwischen der Klasse potentieller Modelle verbundener Theorien, die eines der wesentlichen Elemente eines Theoriekerns sind.

Diese Dissertation befasst sich ausschließlich mit den synchronischen intertheoretischen Relationen. Die strukturalistische Wissenschaftstheorie hat verschiedene intertheoretische Verbindungen definiert, nämlich den *Entailment Link, Determining Link,* die in dieser Dissertation ebenfalls überarbeitete intertheoretische Spezialisierung, intertheoretische Reduktion, intertheoretische Äquivalenz und intertheoretische Approximation. Diese verschiedenen intertheoretischen Verbindungen verbinden die wissenschaftliche Theorie und bilden zwei Arten von Relationen, d. H. Das Theorie-Netz und das Theorie-Holon. Während der Begriff des Theorie-Netzes eine Vorstellung von lokaler oder enger intertheoretischer Relation beinhaltet, beinhaltet der Begriff des Theorie-Holons eine Vorstellung von globaler intertheoretischer Relation.

Die Analyse intertheoretischer Verbindungen in der Kognitionswissenschaft erfolgt in zwei Schritten: Zunächst werden mehrere Theorien formal nach dem Konzept der Theorie-elemente modelliert. Dies sind die Festinger Theorie der kognitiven Dissonanz und ihre Spezialisierung, die Theorie der *forced compliance* Dissonanz aus dem Fachgebiet der Psychologie, die *Computational Neuroscientific Theory* nach Hawkins-Kandel und das McCulloch-Pitts-Modell von Neurons aus dem Gebiet der Neurowissenschaften sowie das Rosenblatt-Perzeptron, das *two-layers feed-*

forward Neuronale Netz, die Delta-Regel und das Hopfield-Netzwerk aus dem Bereich der künstlichen Intelligenz.

Der zweite Schritt ist die Modellierung synchroner intertheoretischer Beziehungen nach dem Konzept des Theorie-Netzes oder nach dem Konzept des Theorie-Holons. Die intertheoretischen Relationen zwischen den obigen Theorien, die modelliert werden, beruhen auf mehreren Forschungen.

Die erste Fallstudie handelt von intertheoretischen Verbindungen im von van Veen et al gemachten Forschung. In diesem Fall ist die Theorie der *forced-compliance* Dissonanz mit der *Computational Neuroscientific Theory* von Hawkins-Kandel verbunden, um die Aktivität von Neuronen in einem Gehirnfeld namens *dorsal Anterior Cingular Cortex* während der Phase der kognitiven Dissonanz zu erklären. Da die Konzepte beider Theorien sehr unterschiedlich sind, gilt für die Modellierung der theoretischen Verbindungen beider Theorien die Definition des *determining links*.

In der zweiten Fallstudie geht es um die intertheoretischen Verbindungen der von Shultz und Lepper entwickelten Simulation, die Konsonanz Modell genannt ist. Dieses Konsonanz Modell basiert auf der Idee: „*dissonance reduction can be viewed as a constraint satisfaction problem*“. (Shultz und Lepper, 1996, S. 220.) Das Modell implementiert das Hopfield-Netzwerk, um ein Subjekt zu simulieren, welches sich in einer Situation oder einer psychologischen Problemstellung befindet und nach der Harmonie seiner Kognition strebt. Die Problemstellung wird durch die klassichen Probleme der kognitiven Dissonanz geschaffen (Shultz und Lepper, 1996, S. 220). In dieser Simulation entspricht die Erhöhung der Konsonanz dem Prozess der Verringerung der Dissonanz oder dem Streben nach Harmonie zwischen den Überzeugungen und Einstellungen des Individuums. Für diesen Fall wird die Definition der *determining links* wieder angewendet, um die

intertheoretischen Verbindungen zwischen dem Hopfield-Netzwerk und der Festinger-Theorie der kognitiven Dissonanz zu modellieren. Anders als im ersten Fall sollte auch das Hopfield-Netzwerk modifiziert werden, damit die Simulation wie erwartet funktioniert.

Der dritte Fall handelt von intertheoretischen Verbindungen einer Simulation der kognitiven Dissonanz, die von van Overwalle und Jordens entwickelt und als *Connectionist* Modell bezeichnet wird. Dieses Modell behandelt einige Aspekte, die vom Konsonanz Modell nicht abgedeckt werden. Die Grundidee dieser Simulation ist inspiriert von der von Cooper und Fazio (1984) vertretenen attributionellen Neuformulierung der Festinger Theorie der kognitiven Dissonanz: Die Reduktion der kognitiven Dissonanz wird von einem rationalen Prozess angetrieben, bei dem das kausale Verständnis von Gedanken, Gefühlen und Verhalten eine wichtige Rolle spielt (van Overwalle und Jordens, 2002, S. 205). Dieses Modell implementiert das *two-layers feed-forward* Neuronale Netz und die Delta-Regel als Trainingsalgorithmus (van Overwalle und Jordens, 2002, S. 206–207), um ein spezifisches Beispiel für ein Kognitionsexperiment zu simulieren, das von Freedman als erstes Paradigma für unzureichende Rechtfertigung bezeichnet wird (1965).

Die Charakterisierung des Modells intertheoretischer Verbindungen für das *Connectionist* Modell erfolgt in mehreren Schritten: Im ersten Schritt soll das Theorieelement des Rosenblatt-Perzeptrons, das Theorieelement des *two layers Feed-Forward* Neuronalen Netzes und das Theorie-Element der Delta-Regel vereinheitlicht werden und ein neues Einheitsmodell von ihnen bauen. Im zweiten Schritt modifizieren wir das neue Einheitsmodell, um es an die Bedingungen des *Connectionist* Modells anzupassen. Der dritte Schritt besteht darin, die Theorie der forced-compliance Dissonanz zu modifizieren, indem einige Konzepte gemäß der Idee von Cooper und Fazio hinzugefügt werden. Im letzten Schritt sollen die intertheoretischen

Verbindungen zwischen der modifizierten Festinger-Theorie der kognitiven Dissonanz und dem modifizierten Einheitsmodell des Rosenblatt-Perzeptrons, dem *two layers feed-forward* Neuronalen Netz und der Delta-Regel charakterisiert werden. Diese Modellierung besteht aus verschiedenen intertheoretischen Verbindungen, wie z. B.: *Determining links*, Spezialisierung und Reduktion.

Die letzte intertheoretische Verbindung zu modellieren ist die intertheoretische Verbindung des McCulloch-Pitts-Neurons und des Rosenblatt-Perzeptrons. Dieses Modell zeigt, wie zwei in zwei verschiedenen Disziplinen entwickelte Theorien sich in Beziehung setzen können, weil eine davon die andere verallgemeinert. In diesem Fall ist das später in der künstlichen Intelligenz entwickelte Rosenblatt-Perzeptron eine Weiterentwicklung des zuerst in den Neurowissenschaften entwickelten McCulloch-Pitts-Neurons. Aus der synchronischen Perspektive ist die intertheoretische Verbindung beider Theorien eine intertheoretische Reduktion, bei der das McCulloch-Pitts-Neuron vom Rosenblatt-Perzeptron reduziert wird. Da beide Theorien nahe beieinander liegen, handelt es sich bei diesem Fall nicht um die globale intertheoretische Relation, sondern um die lokale intertheoretische Relation. Daher müssen wir lokal intendierte Anwendungen nicht besprechen.

Diese Forschung liefert mehrere Ergebnisse, die nicht nur in der Philosophie, sondern auch in Studien in interdisziplinären Bereichen, insbesondere der Kognitionswissenschaft, einige Beiträge leisten: Erstens schlägt diese Dissertation eine kleine Überarbeitung der Definition der intertheoretischen Spezialisierung und einer Weiterentwicklung für den Begriff des Theorie-Holons, d.h. des V-Musters und der V-strategie vor. Das V-Muster und die V-Strategie wurden als Werkzeuge entwickelt, um mehrere Theorie-Elemente zu kombinieren und daraus ein neues Einheitstheorie-Element zu erstellen. Dieses V-Muster und diese V-

Strategie haben einen praktischen Zweck, der auf dem Konzept des Theorie-Holons und seinem empirischen Klaims basiert, nämlich wie man mehrere Theorien in einer globalen intertheoretischen Relation kombiniert und seine lokale intendierte Anwendung bestimmt, indem nur dyadische Verbindungen implementiert werden.

Zweitens, da es sich bei allen untersuchten Fällen um eine intertheoretische Reduktion handelt, können wir das strukturalistischen Modell mit dem generalisierten Nagel-Schaffner Modell (das GNS-Modell) vergleichen. Die Unterschiede von beiden lauten wie folgt: (1) Der Hauptunterschied liegt in der strukturalistischen Forderung, dass die wissenschaftlichen Theorien in der Mengenlehre modelliert werden müssen, um ihre innere logische Struktur zu modellieren. Dieser Ansatz erfordert eine kompliziertere Modellierung als das GNS-Modell, bietet jedoch gleichzeitig mehr Details und eine genauere Analyse. (2) Der zweite Hauptunterschied besteht darin, dass nach Ansicht der Strukturalisten die intertheoretische Reduktion eher epistemologisch als ontologisch ist, obwohl sie eine empirische Basis haben sollte - charakterisiert durch die partiellen Potentialmodelle. Es bezieht sich eher auf die Struktur der Theorien als auf die Realität. (3) Der dritte Hauptunterschied besteht darin, dass in der strukturalistischen Theorie der Wissenschaften der Strukturalisten die intertheoretische Reduktion nur eine von mehreren anderen intertheoretischen Verbindungen (Verknüpfungen) ist und die Strukturalisten einige von ihnen bereits formal charakterisiert haben. (4) Es gibt kein als solches verallgemeinertes Modell der intertheoretischen Reduktion für die Strukturalisten. Die Strukturalisten haben eine formale Definition als ein Werkzeug definiert, wie eine intertheoretische Reduktion von einer wissenschaftlichen Theorie auf einer anderen Theorie modelliert werden kann, aber die Strukturalisten haben nicht die Absicht, ein allgemeines Reduktionsmuster für wissenschaftliche Praktiken zu formalisieren. (5) Der letzte Unterschied soll zeigen, dass die

intertheoretische Reduktion etwas mit den erklärten Phänomenen zu tun hat, das strukturalistische Modell verwendet die *interpreting links*, die die Potentialmodelle eines Theorie-Elements mit dem partiellen Potentialmodell der anderen Theorie-Elements verbinden. Das GNS-Modell verwendet einige Verbesserungen, die von Dizadji-Bahmani, F., Frigg, R. und Hartmann S. (2009) in ihrer Arbeit „*Who's afraid of Nagelian Reduction?*" Und von van Riel, Raphael (2011) in seiner Arbeit vorgeschlagen wurden „*Nagelian Reduction beyond the Nagel Model*".

Drittens zeigt diese Dissertation meine einzigartige Position in Bezug auf den Begriff der Einheit der Wissenschaft in der Wissenschaftstheorie. Im Allgemeinen gibt es zwei entgegengesetzte Positionen in Bezug auf den Begriff der Einheit der Wissenschaft in der Wissenschaftstheorie, nämlich die Stanford-Schule, die den Begriff der Einheit der Wissenschaft auf der Grundlage der Metaphysik der Wissenschaft aufgibt, und andere Philosophen, die immer noch einen Begriff der Einheit der Wissenschaft vertreten. Die Position stimmt mehr mit der Stanford-Schule überein, obwohl diese Dissertation sieht, dass die wissenschaftlichen Theorien auf der Grundlage der wissenschaftlichen Praxis miteinander verbunden sind. Diese Position ist sehr ähnlich zum integrativen Pluralismus.

Viertens, in interdisziplinären Forschungen, insbesondere in der Kognitionswissenschaft, schlägt meine Dissertation einen Ansatz vor, wie wir intertheoretische Verbindungen zwischen wissenschaftlichen Theorien modellieren und analysieren oder verschiedene wissenschaftliche Theorien kombinieren können.

Acknowledgment

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dr. em. Dr. h.c. mult. Carlos Ulises Moulines for the continuous support of my Dr. Phil. study and related research for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Dr. Phil. study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Dr. Bernhard Lauth, and Prof. Dr. Volker Tresp, for their readiness to be my supervisor in my disputation.

My sincere thanks also go to Prof. Dr. Bernhard Lauth, Dr. Connor Mayo-Wilson (University of Washington), and Prof. Dr. Gregory Wheeler (University of Frankfurt) and Victoria Schöffel, who helped me to find the topic and some cases of my research. Without their precious support, it would not be possible to start this research.

I thank Prof. Dr. Stephan Hartmann (MCMP-LMU), Dr.Phil. habil. Alexander Reutlinger (MCMP-LMU), and Dr. Karolina Kryzanowska (University of Amsterdam/MCMP-LMU) for ideas and inputs in the lectures or for some papers and presentation that I made to develop my ideas for my dissertation. I thank Prof. Dr. Volker Tresp (LMU), who helped me to learn artificial neural networks and machine learning. My thank goes to some professors in the department of the neurobiology of the LMU who helped me in my confusion when I began to learn some topics in neuroscience. I am very sorry that I can not remember their name anymore.

My dissertation heavily relies on several researches conducted and models built by others. Therefore, I would like to thank Prof. Dr. Reiner Westermann (University of Greifwald), Prof. Dr. Thomas R. Shultz (McGill

University), Prof. Dr. Mark R. Lepper (Stanford University), Dr. John Bickle (Mississippi State University), Prof. Dr. Frank van Overwalle (University of Brussel), and Dr. Vincent van Veen (Diablo Valley College) for some contacts and helping me to understand the materials.

I thank PD. Dr. Dr. Thomas Brückner (LMU), Prof. Dr. Holger Andreas (University of British Columbia Okanagan), Dr. Lena Höfer, and Dr. Marek Polansky for helping me to learn the structuralist theory of science and all logics needed during my magister study.

I thank Dr. Tilman Massey, Jessica Intane, and Omar Rodriguez Carrasquillo for supporting me with some materials I need for my dissertation.

My thanks go to Dr. Neil Dewar (MCMP-LMU), John Peitz, Dr. Daniel Molnar, Irwan Tjulianto, and Jessy Siswanto for proofreading my dissertation as well.

Last but not least, I would like to thank my family and friends. I thank my mother for supporting me spiritually throughout writing this thesis and in my life in general. I thank my friends in MRII Berlin, Hamburg, Munich, Bern, and Stockholm for cheering and supporting me spiritually while writing my thesis.

Contents

List of Tables

List of Figures

Chapter 1
The Problem and Its Context: Background and Relevance

1.1. Interdisciplinarity: Its Brief History, Concept, and Aspect(s) Related to This Project

'Interdisciplinarity' is one of the most important words to describe a trend in the scientific world today. As Palmer has put it, "interdisciplinarity has become a topic of wide interest, penetrating the sciences, social sciences, and humanities. Many researchers practice it, and others study it. Scholars in the emergent area of knowledge studies have made many observations that call attention to the importance of interdisciplinary inquiry for the advancement of knowledge. For example, they have claimed that path-breaking ideas usually come from the cross-disciplinary investigation (Turner, 1991) and that disciplinary boundaries are the fault line that conceals future scientific revolutions (Fuller, 1988)." (Palmer, 1996, p. 30). This trend of interdisciplinarity began in the 20th century and has grown stronger today. Nevertheless, what is 'interdisciplinarity'? Why has this trend become so popular in the sciences? Why is it important to understand? This dissertation will start to answer these questions by giving a simple definition of the word 'interdisciplinary,' telling a brief story of the development of science, and delivering in-depth explanations of this trend.

The Meaning of 'Interdisciplinarity' and 'Interdisciplinary.' The word 'interdisciplinary' is a compound adjective word consisting of a prefix 'inter' and the word 'disciplinary.' The prefix 'inter' means "between; among" or "mutually; reciprocally" (Oxford Dictionary of English). The word 'disciplinary' means "concerning or enforcing a discipline," and the word 'discipline' means "a branch of knowledge" (Oxford Dictionary of English). The Oxford Dictionary of English gives a simple meaning for

‘interdisciplinary’ as ‘relation between more than one branch of knowledge’ and for ‘interdisciplinarity’ as ‘the quality or fact of involving or drawing on two or more branches of knowledge.’ Before we try to give a deeper and more precise understanding of what ‘interdisciplinary’ and ‘interdisciplinarity’ really mean, it would be beneficial to get first an intuitive idea by examining its historical background in the history of science.

A Brief History of Interdisciplinarity. Concerning the idea of interdisciplinarity, the history of science can be divided into three periods: the period of pre-disciplinarity, the period of shaping disciplinarity, and the period of developing interdisciplinarity. In the pre-disciplinary period, the search for knowledge seemed to be interdisciplinary because there were no specialized branches of knowledge like in current time. Several philosophers in ancient Greece and Rome, like Aristotle and the stoic philosophers, tried to categorize human knowledge and to understand how that knowledge was best gained and ordered. “Aristotle differentiated ‘*scientia*’ (episteme) as the knowledge about causes and reasons from mere opinions (*doxa*) that are often subjective, and from technology (*techne*) and the arts (*ars*) as the knowledge requisite to create or construct. In this classification, only scientific knowledge (*scientia*) can claim to be universally valid. Science is, distinct from practical orientation, a theoretically oriented activity. Theoretical knowledge is gained by observation and contemplation and comprises three areas (or disciplines in the modern sense): mathematics, physics, and (first) philosophy. Mathematics consists of geometry, arithmetic optics, and harmonics. Physics is the knowledge of the material world and all forms of life (i.e., today’s biology). Philosophy includes knowledge of the cosmos and theology. The Roman Stoa (c. 300 BC) subsequently developed a classification of knowledge in opposition to Aristotle’s that included practical knowledge and distinguished logics, physics, and ethics. Subsequently, Aristotelian and Stoic classifications overlapped and merged with the medieval concepts of the ‘*artes liberales*’ that constituted what was then

considered the comprehensive system of knowledge: grammar, rhetoric, logic, arithmetic, music, geometry, and astronomy." (Weingart, 2010, p. 3).

After that, the classification and the theory of science and knowledge production continued to be developed. For example: "Bacon differentiated, on the one hand, between natural and civil history dealing with works of nature and man, respectively. On the other hand, the sciences were distinguished into theology and philosophy, and philosophy into a doctrine of the deity (natural theology), a doctrine of nature and a doctrine of man. … The hierarchy of faculties as organizational structures of the universities, institutionalized since the Middle Ages, placed philosophy at the bottom, and medicine, law, and theology above it. Although this hierarchy of the faculties had also represented a classification of knowledge, it subsequently lost acceptance. At the end of the eighteenth century, the notion of a 'lower' faculty and 'higher faculties' counted as past." (Weingart, 2010, pp. 4–5).

The epoch of shaping disciplinarity already began in the time of Bacon. However, at that time, "these disciplines did not have a social function of their own but only served as repositories of certified knowledge. Disciplines were relatively unimportant until the end of the eighteenth century (Stichweh 1984, pp. 14–15)." (Weingart, 2010, p. 4). At the end of the eighteenth century, the notion of discipline took over the role of the hierarchy. The most important reason for this was the growing pressure that data collection had on the disciplines. It caused problems of overload and integration (Weingart, 2010, p. 4).

To solve this problem, the scholarly activities were differentiated into disciplines through two developments, namely increasing abstractions and the number of new subject matters. Weingart writes: "One was increasing abstraction, for example, through the mathematical conceptualization of objects. It means that science to a decreasing degree gained its information about the world directly from its environment. … A growing stock of concepts, theories, and instruments mediated the experiences gathered, i.e., experience

was no longer grasped immediately but rather constructed on the level of the concept. At the same time, and this is the second development, the modern scientific mode of gaining knowledge was expanded to new subject matters." (Weingart, 2010, pp. 5–6). This development brought some other results, namely: (1) The birth of specialization, (2) the self-referential specialized communication among the scholars, that caused a division between specialists and laypersons, (3) the loss of the unity of science, (4) a fundamental change in the orientation of scientists from becoming knowledgeable in several fields of science to discovering new phenomena and explanations, (5) organizing researches on the basis of a division of labor into numerous highly specialized activities, and (6) switching roles between academic and university: the academies became the institutional place for the collection and conservation of knowledge, while the university produced and disseminated new knowledge (Weingart, 2010, pp. 6–7).

According to Weingart, a discipline is formed and developed through self-referential communication, which is 'closed' towards its environment; the evaluation of relevance and quality of research is limited to the members of the respective disciplinary community (Weingard, 2010, p. 8). Therefore, every discipline has a social identity and a factual identity. Discussing both identities, Weingart writes: "Their social identity is constituted by the rules of membership, i.e., teaching, examinations, certificates, careers, the attribution of reputation, and, thus, the formation of a hierarchical social structure. Their factual identity is constituted by the contents of the communication. It concerns the delineation of a subject matter, a common set of problems and theories, concepts and specific methods to study it, the criteria of quality of achievement which are the basis for the evaluation and attribution of reputation by peer review" (Weingart, 2010, p. 8).

Institutionalization of disciplines occurs not only in university faculties but also in scholarly associations. These scholarly associations have functions not only for the internal aspects of the disciplines via coordinating

the communication process by staging conferences and running disciplinary journals, but also for their economic, political, and social environments via representing the interests of the disciplinary communities in various ways, such as the certification of disciplinary training and formal accreditation as attempts to secure a monopoly for a specific sector of the professional or semiprofessional job market (Weingart, 2010, pp. 8–9). Although the disciplines have a significant degree of autonomy in determining their development, they depend on external resources that are distributed by research councils and foundations – government departments and industry – according to their priorities and political or economic goals (Weingart, 2010, p. 9).

The epoch of interdisciplinarity began with the uneasiness about the loss of the unity of science, which brought back a call for reunification or interdisciplinarity. According to Weingart (2010), several successive occasions mark the coming of this new epoch (p. 12): First, in the 1930s the first 'unity of science' movement was initiated by philosophers of science and natural scientists, especially those of the Vienna Circle, but this movement had no impact. Second, in the late 1960s and the 1970s in the context of debates about technology gaps, technology forecasting, and protection of the environment, the Organisation for Economic Cooperation and Development (OECD) triggered a new debate on interdisciplinarity (Apostel et al., 1972). Erich Jantsch's term 'transdisciplinarity' from that publication was used by Gibbons et al. (1994) to diagnose the emergence of a new mode of knowledge production. Gibbon named it 'mode 2.' His thesis states that the traditional disciplinary 'mode 1' of knowledge production has given way to a new transdisciplinary mode of knowledge production. Since that time, there were many animated discussions among analysts and the mobilization of conflicting evidence. Together with a series of similar pronouncements of a fundamental change in knowledge production, these analyses beg the

question of whether they genuinely signal the advent of a new order of knowledge formation or if they only describe surface phenomena (Funtowicz and Ravetz, 1993; Ziman, 1994). The difference between 'interdisciplinary' and 'transdisciplinary' will be discussed later in this chapter – for our purpose here, it is sufficient for us to look at the history of interdisciplinarity and some essential, related terms have been discussed.

The claims of change – insofar as they are relevant here – can be summarized as follows: "the university has lost its monopoly as the institution of knowledge production since many other organizations are also performing that function. Transitory networks and contexts are formed which replace traditional disciplines. Knowledge production outside disciplines is no longer the search for basic laws (fundamental research) but takes place in contexts of application (Funtowicz and Ravetz, 1993, p. 121; Gibbons et al., 1994, p. 4). Disciplines are no longer the crucial frames of orientation for the delineation of subject matters and the formulation of research problems. Research is, instead, characterized by transdisciplinarity: solutions to problems appear in contexts of application, and research results are no longer communicated in journals. The criteria of quality are no longer determined by disciplines alone, but additional criteria, social, political, and economical, are applied to determine quality (Funtowicz and Ravetz, 1993, p. 90; Gibbons et al., 1994, p. 8)" (Weingart, 2010, p. 12).

There are two historical and sociological reasons for the emergence of inter- and transdisciplinary structures that would replace traditional disciplines. As Weingart writes: "First, with the continuously growing number of specialties (i.e., research fields below the level of disciplines) the probability increases that, due to the proximity of such fields, new re-combinations will occur which will result in new 'interdisciplinary' research fields. ...Second, inter- and transdisciplinary research fields are promoted by funding agencies in the interest of directing research to politically desired

goals. This process is conditioned by the fact that the 'externally' defined subject matters, research problems, and values or interests can trigger sustained research" (Weingart, 2010, p. 12, see also Jungert, 2013, p. 10).

The Concept of Interdisciplinarity. It is not easy to give an exact definition of the term interdisciplinarity. In several works discussing interdisciplinarity, there exist different ways to define interdisciplinarity. The first way is defining interdisciplinarity by the goals of the cooperation. There are various ways of applying multiple disciplines at the same time for the same purpose. And not all of them deserve to be called 'interdisciplinary.' A relationship between management sciences and philosophy cannot be called 'interdisciplinary' just because a university applies several management theories in planning to start a philosophical department. To be called 'interdisciplinary,' these various disciplines must be related to each other in such a way that in their activities their relationship has a purpose of gaining knowledge (Voigt, 2013, p. 32). In search of knowledge, the activities, such as changing, limiting, or expanding the object-fields and modifying, newly developing, or giving up the methods beyond the constraints of the disciplines, are the elements of interdisciplinary cooperation (Voigt, 2013, p. 32). A comprehensive definition is given by the Organization for Economic Cooperation and Development: "Interdisciplinary—An adjective describing the interaction among two or more different disciplines. This interaction may range from a simple communication of ideas to the mutual integration of organizing concepts, methodology, procedures, epistemology, terminology, data, and organization of research and education in a fairly large field. An interdisciplinary group consists of persons trained in different fields of knowledge (disciplines) with different concepts, methods, and data and terms organized into a common effort on a common problem with continuous intercommunication among the participants from the different disciplines" (OECD, 1972, pp. 25–26). To find out whether cooperation between disciplines is interdisciplinary or not, Birnbaum (1977) developed a set of

indicators to determine the extent to which a project meets the criteria for interdisciplinary research. These indicators are: "(1) different bodies of knowledge are represented in the research group, (2) group members use different problem-solving approaches, (3) members of the group perform different roles in solving problems, (4) members of the group work on a common problem, (5) the group is responsible for the final product, (6) the group shares common facilities, (7) the nature of the problem determines the selection of group members, and (8) members are influenced by how others perform their tasks" (in Lattuca, 2001, p. 13).

Several Other Related/Similar Terms. In many books about interdisciplinarity, two other terms refer to a similar kind of relation, where various disciplines stand together, namely multidisciplinarity and transdisciplinarity.[1] Therefore, it is important here to explain the differences between them. Multidisciplinarity refers to several different disciplines speaking about a common or similar theme without trying to build any structured cooperation or synthesis among their results. The most striking difference from pure disciplinary research is that in multidisciplinary research there is a minimal knowledge about the relevant research from the other fields (Jungert, 2013, p. 2). Klein says that juxtaposing, sequencing and coordinating are the characteristics of multidisciplinarity (Klein, 2010, p. 16). The concept of transdisciplinarity means more intensive cooperation between

[1] Jungert (2013) gives several other words related to the notion of interdisciplinarity. They are multi-, pluri-, cross-, inter-, and transdisciplinarity. Here we discuss only multi-, inter-, and transdisciplinarity because these three represent a full relationship among various disciplines. The word 'pluridisciplinarity' is rather a synonym for multidisciplinarity. It describes a relationship similar to multidisciplinarity with the extra notion "to enhance the relationship between them" (Jantsch, 1970 in Jungert, 2013, p. 2). The word 'crossdisciplinarity' refers to the following phenomenon: "The axiomatics of one discipline are imposed upon other disciplines at the same hierarchical level, thereby creating a rigid polarization across disciplines toward a disciplinary axiomatics" (Jantsch, 1970 in Jungert, 2013, p. 3). In this cross-disciplinarity methods and research programs of another discipline are taken over as its subject. Its goal is neither a fusion of given disciplines nor a molding of new disciplines (Jungert, 2013, p. 3), but "to solve important and urgent problems that cannot be defined and solved from the perspective of any one of the existing disciplines" (Kockelmans, 1979, p. 82).

the disciplines that finally lead to the crossing and merging of different disciplines. This cooperation leads to a continuous system of scientific, systematic change that changes the technical and disciplinary orientations (Sukopp, 2013, pp. 23–24). Klein describes the characteristics of a transdisciplinarity as transcending, transgressing, and transforming (Klein, 2010, p. 16). By contrast, the concept of interdisciplinarity was defined by Heckhausen: "*Die Rede von der ‚interdisziplinären Forschung' besagt gewöhnlich nicht mehr, als dass einige Wissenschaftler, die verschiedenen Fächern gehören, zusammen an einem Problem arbeiten, das so allgemein, alltagsnach oder fachfremd betitelt ist, daß noch kein Vertreter der beteiligten Fächer bereits das Problem unter den Aspekten seiner eigenen Fachlichkeit eingegrenzt und definiert hätte*. [In English: The talk of 'interdisciplinary research' usually just means that some scientists who belong to different disciplines work together on a problem that is so general, every day or unfamiliar that no representative of the specialties involved has the problem delimited and defined from the point of view of his own specialty.]" (Heckhausen, 1987, p. 129). This leads to working or researching together in cooperative scientific action. As for the characteristics of interdisciplinarity, Klein (2010) mentions integrating, interacting, linking, focusing and blending. (P. 16). These differences can be summarized in the following table:

Relation	Multidisciplinarity	Interdisciplinarity	Transdisciplinarity
Characteristics	juxtaposing, sequencing, and coordinating	integrating, interacting, linking, focusing, and blending.	transcending, transgressing, and transforming
Cooperation	no cooperation	cooperation	cooperation

Table 1.1. The differences between multidisciplinarity, interdisciplinarity, and transdisciplinarity regarding their characteristics and the cooperation between their discipline-elements. (Adapted from Klein, 2010, p.16)

Aspects of Interdisciplinary Studies or Researches Regarding This Project. According to Jungert, there are five aspects of a discipline that play an important role in examining the characteristics of interdisciplinary relations: *Gegenstände* (objects of investigation), *Methoden* (methods), *Probleme* (problems), *theoretisches Integrationsniveau* (level of theoretical integration), and *Personen/Institutionen* (persons/institutions) (Jungert, 2013, pp. 7–9). These five aspects must be present to form an interdisciplinary relationship. However, there are different degrees of each among various interdisciplinary researches. This dissertation will focus solely on one of the biggest challenges in interdisciplinary researches related to the issues of theoretical integration; how can we connect scientific theories from various disciplines such that the intertheoretical relation built is fruitful? This work aims to suggest a formal approach to build a model of how several theories from various disciplines are related to each other in interdisciplinary research.

Cognitive Science as A Case Study of Intertheoretical Relations. The subject of interdisciplinarity is both broad and vast. This dissertation will only focus on one interdisciplinary field, namely cognitive science. The reasons for choosing cognitive science are as follows: (1) Cognitive science fulfills the criteria of interdisciplinary studies. For studying the mind, cognitive science integrates many fields of science. Furthermore, it is not merely that several different disciplines speak about the same or a similar theme without any structured cooperation or trying to build a synthesis of their results – cognitive science is not merely multidisciplinary. (2) There are many fruitful researches in cognitive science. Intuitively, we are right in hoping that we can learn more from something successful, rather than from something unsuccessful or not yet successful. (3) Cognitive science is not only interdisciplinary but also a multicategory discipline. The term ‘categories’ here refers to our classification of the various scientific disciplines into three categories that we normally use, i.e., natural sciences,

social sciences, and humanities. The disciplines in cognitive science do not only belong to natural sciences but also the social sciences and the humanities.

However, cognitive science is a vast interdisciplinary field, that contains at least six disciplines, hundreds of theories, and research programs – and it seems that these numbers are still increasing. Therefore, there shall be a limitation of the number of cases discussed. This dissertation will build and analyze formal models for the intertheoretical relation between cognitive dissonance and the corresponding computational neuroscientific theory in the functionality of *dorsal Anterior Cingulate Cortex* (*dACC*), between two models of neurons, and two simulations of the theory of cognitive dissonance. The first simulation is the consonance model built by Thomas R. Shultz and Mark R. Lepper and the second is the adaptive connectionist model built by Frank van Overwalle and Karen Jordens. The purpose of modeling several models of intertheoretical relations is to show how this approach is also applicable to many other synchronic intertheoretical relations. Synchronic intertheoretical relations occur in those cases where scientists connect several theories assumed existing together for their research. The complete description of these combined theories and their place in cognitive science will be discussed in Chapters 4–8.

1.2. State of the Art of the Problem: Studies about Intertheoretical Relations in Interdisciplinary Fields

Although the topic of intertheoretical relations has recently become more and more interesting for philosophers of science, the discussions are mostly dominated by the discussions about reduction or intertheoretical reduction.[2] This fact is understandable for the two following reasons: an ontological reason and a historical reason.

[2] In several renowned encyclopedias of philosophy, a subject of intertheoretical relation cannot be found in their table of contents, whereas a subject of reduction or intertheoretical reduction can. These encyclopedias are the following: the Stanford Encyclopedia of

Historical reason. Although many aspects of science have been thought over since the Greek philosophers, modern philosophy of science began with Ernst Mach's program and logical positivism. They systematically consolidated this heritage into a brand new approach to philosophy and science by implementing modern logic and engaging in the conceptual foundations of empirical science. The goals of their program were (1) to create new foundations for physics with a strong consideration of the results of the physiology of the senses, (2) to recover the unity of all sciences, and (3) to eradicate the metaphysical speculations from the field of science for good (Moulines, 2008, p. 26). Thus, the topic of reduction became one of the most critical issues in the agenda of modern philosophy science. A monumental work from one of those thinkers, *Der logische Aufbau der Welt* by Rudolf Carnap (1928), tries to reduce physics and the other sciences to elementary psychology. Carnap's other work, *Unity of Science*, is an attempt to build the unity of science from the language of physics by reduction: "… science is a unity, [such] that all empirical statements can be expressed in a single language, all states of affairs are of one kind and are known by the same method" (Carnap, 1934, p. 32). Although the reduction is not the only epistemological proposal for the unity of science today, this topic is still an important and relevant topic with a most extensive historical background in philosophy of science.

Ontological reason. To obtain a richer explanation of a specific phenomenon x, it is evident that we can refer to more basic phenomena, which constitute x. Therefore, the concept of reduction is a significant component in the discussion about intertheoretical relations. According to the Stanford Encyclopedia of Philosophy, the reduction can be understood as follows:

Philosophy, The Internet Encyclopedia of Philosophy, the Routledge Encyclopedia of Philosophy, and Kaldis, Byron Ed. (2013): The Encyclopedia of Philosophy and Social Science. None of these have an entry about intertheoretical relations, but all of them have an entry about reduction.

> "The English verb 'reduce,' derives from the Latin '*reducere*,' whose literal meaning 'to bring back,' informs its metaphorical use in philosophy. If one asserts that the mental reduces to the physical, that heat reduces to kinetic molecular energy, or that one theory reduces to another theory, one implies that in some relevant sense the reduced theory can be *brought back* to the reducing theory, the mental can be *brought back* to the physical, or heat can be *brought back* to molecular kinetic energy. The term 'reduction' as used in philosophy expresses the idea that if an entity *x* reduces to an entity *y*, then *y* is in a sense *prior to x*, is *more basic than x*, is such that *x fully depends upon* it or is *constituted by it.* Saying that *x* reduces to *y* typically implies that *x* is *nothing more than y* or *nothing over and above y.*
>
> ... 'Reduction' is a term of natural language. Building upon its common metaphoric meaning, philosophers use it to designate relations of particular philosophical importance in many closely related fields, especially in the philosophy of science, the philosophy of mind, and metaphysics.
>
> The notion of *scientific* reduction as used in contemporary analytic philosophy differs from conceptions of reduction according to which we learn about the instantiation of reduction relations on a purely *a priori* basis from basic religious, metaphysical or epistemological principles. 'Scientific reduction' applies to reductionist claims supposedly justified by scientific evidence and the success of science." (van Riel and van Gulick, 2014).

It is important to note that this concept of reduction is not to be confused with reductionism, which is also a popular term or school in philosophy of science. "Reductionism is the adoption of reduction as the global ideal of the unified structure of scientific knowledge and a measure of its progress." (Cart, 2013). People can agree with the existence of reduction as an intertheoretical relation, without agreeing with reductionism.

Because of its importance, there are attempts to conceptualize the concept of reduction or, respectively, the concept of scientific reduction. One of the popular approaches is the generalized Nagel-Schaffner (GNS) model of reduction, outlined by Ernst Nagel in his *Structure of Science* and improved by Schaffner and others. In Chapter 8 the GNS model of reduction will be compared to the structuralist models of intertheoretical reduction. In this dissertation, the structuralist theory of science will be applied to modeling intertheoretical reduction in some existing researches and analyzing them to understand how intertheoretical reduction works.

At this point, it is important to be clarified that this dissertation has no agenda to develop or to support any idea of reductionism. However, the purpose of this dissertation in this matter is to propose an approach to evaluate the modern notion of intertheoretical reduction and the notion of reductionism by applying the structuralist metatheory of science.

Related to intertheoretical reduction, another issue to consider is the idea of the unity of science. In the birth of modern philosophy of science, recovering unity of science was set as one of its goals, but now there are many discussions to evaluate this goal – or respectively the idea. The reasons are: (1) there is no sign of its realization until now. Furthermore, (2) the idea of the unity of science, especially in its original version, is merely incompatible and implausible concerning the real scientific practices and the real development of science (Dupré, 1995). Members of the Stanford School, such as Dupré, Ian Hacking, Peter Galison, Patrick Suppes, and Nancy Cartwright, have launched attacks against universalism and uniformity both in the methodological and the metaphysical sense: “Disunity appears characterized by three pluralistic theses: against essentialism, there is always a plurality of classifications of reality into kinds; against reductionism, there exists equal reality and causal efficacy of systems at different levels of description, that is, the micro level is not causally complete, leaving room for downward causation; and against epistemological monism, there is no single

methodology that supports a single criterion of scientificity, nor a universal domain of its applicability, only a plurality of epistemic and non-epistemic virtues" (Jordi, 2013, about Dupré). On the other side, many philosophers, such as Hempel, Nagel, Friedman, Kitcher, Cat, Klein, Putnam, etc., still think that the unity of science is an important idea. They propose many ideas to evaluate and sharpen the concept in many ways by asking the following questions: Is the unity of science ontological or epistemological? How can the unity of science be achieved – by which method, under which requirements, etc.? Of course, my position can easily be ascertained from the title of my dissertation. This dissertation assumes that there is so-called unity of science, but the kind and the way still need to be clarified. Hopefully, this dissertation can provide some clues to make the discussion about the unity or disunity of science more fruitful (in Chapter 8) by giving some concrete examples from the real scientific practice in interdisciplinary fields, especially in cognitive science (in Chapters 5–7).

1.3. Studies on Intertheoretical Connections in Cognitive Science

As an interdisciplinary field, cognitive science also becomes an exciting field for discussions about intertheoretical relations. In the discussion about the reduction mentioned above, the relationship between the mind and the body, which is one of the most central topics in cognitive science, often becomes an object of debate. During the rapid development of cognitive science, there are four views about intertheoretical connections in cognitive science according to Ezquerro and Manrique (2004), namely: (1) the classical view, (2) the connectionist revision, (3) the pragmatist approach, and (4) the reductionist approach. This categorization is based on the position of these authors with respect to the notion of the privileged level. The privileged level is understood here as a level (or a discipline) "at which all the different disciplines come to converge" (Ezquerro and Manrique, 2004, p.61): "Research in Cognitive Science has often assumed the existence of a

privileged level at which all the different disciplines come to converge. Computational theories were the first ones to offer themselves as such a level. The equation 'cognition = information processing,' on which the cognitive revolution was founded, seemed to find its natural place in the technical and mathematical developments provided by those theories. The possibility of obtaining a system of computational mechanisms that accounted for the totality of cognitive phenomena offered the promise of a 'unified theory of cognition' (Newell, 1990)." The classical view and the connectionist revision agree with the existence of such a privileged level, without being reductionist. The classical view holds the view that "to explain a particular mental phenomenon … required giving the right computational account between the right kinds of representations, which were conceived as a symbol system" (Ezquerro and Manrique, 2004, p. 65), whereas the connectionist approach does not use a symbol system or manipulations of symbols but an artificial neural network. However, the proponents of connectionism claim that " their systems offer a real possibility to bridge neuroscientific, computational, and intentional descriptions" (Ezquerro and Manrique, 2004, p. 68). The pragmatist approach denies the existence of a privileged level because they see that "the different ways of formulating levels depend just on our different approaches to the phenomena we want to study, and this, …, depends on pragmatic considerations" (Ezquerro and Manrique, 2004, p. 81). Furthermore, the reductionist approach, as proposed by Bickle, agrees with the existence of a privileged level and tries to give a more detailed and smooth intertheoretical reduction using the set-theoretical approach characteristic of the structuralist theory of science. Chapter 8 will also respond to this discussion about the intertheoretical reduction in cognitive science and show the position and the contributions of this research concretely.

1.4. Why the Structuralist Approach?

From the description of the current state of the research on intertheoretical relations both in the philosophy of science and cognitive science above some points can be concluded as follows: (1) The discussions about intertheoretical relations are dominated by the idea of reduction as if there were no other intertheoretical relations[3] (2) Most examples presented in the discussion of intertheoretical relations are reasonable speculative sketches about intertheoretical relations – the theories are real, but the models of intertheoretical relations presented have just a minimal relation – if not without any relation at all – to scientific practice. Therefore, there is no way to verify those models, and they only give small contributions to interdisciplinary practices. (3) There may be models of intertheoretical relations that are based on real scientific practice besides the models based on speculative sketches. However, they are not detailed enough to give significant contributions to scientific practices, and (as far as the author can see) none is a model of interdisciplinary intertheoretical relations.

Because of these points, this research focuses on building a model of intertheoretical relations for a specific interdisciplinary field and take cognitive science as a real case. The models are based on concrete examples in cognitive science, namely the explanation or simulation of a psychological property by the network in the brain or the artificial neural network. It is generally assumed that our mental properties or faculties are connected to our brain, and several scientific research about these connections have been done in cognitive science. The implementation of artificial neural networks comes

[3] Of course, I do not mean to say that all philosophers of science know only intertheoretical reduction. Many philosophers surely know that there are different kinds of intertheoretical relations besides intertheoretical reduction. Structuralists know several types of intertheoretical relations, and intertheoretical reduction is just among them (see Chapter 2). We may also look at Schaffner's revision of Nagel's notion of reduction called the Generalized Nagel-Schaffner (GNS) model of reduction.

into the playground because it is currently still impossible for us to understand the networks of the brain entirely.

The structuralist theory of science is chosen as the underlying theory for this project because of the following reasons: (1) To trace how scientists add some constraints or assumptions to the theories and how those constraints or assumptions affect the theories, it requires a theory that can help for building a detailed model of the combined theories and their relations. The structuralist metatheory of science provides tools for analyzing and building models of scientific theories by identifying their important components. (2) The structuralist theory of science implements the formal approach that enables us to reach high clarity and consistency in modeling and analysis. It is crucial because some terms from various disciplines often have a different meaning or reference, although they use the same word. To understand intertheoretical connections between theories from various disciplines, it is important to identify precisely which parts of the theories in question are interconnected and how they are connected. The structuralist theory of science provides powerful tools by modeling the theory in several classes of terms or their relations, i.e., potential models, actual models, etc. With these classes, it becomes possible to identify not only the inner structure of the connected theories but also the connected parts of the theories and the properties of the connections. Modeling individual theories in question will be presented in Chapter 3, whereas modeling intertheoretical connections between them will be in Chapters 5–7. (3) The structuralist theory of science has high flexibility for modeling because it implements set theory instead of first-order predicate logic. (4) The structuralist theory of science has already identified and characterized several kinds of intertheoretical connections. It will be of a great help in analyzing the intertheoretical relationships, which no other approach in the philosophy of science can offer.

By applying the structuralist theory of science, the research will be conducted as follows. First, several formal models of the theory-elements that

will be combined are presented in Chapter 3. After that, some preparations for the modeling of intertheoretical connections in interdisciplinary fields should be done in Chapter 4, i.e., explaining some formal tools and revisions that will be helpful for the modeling and a short overview of cognitive science related to our investigation. Chapters 5–7 will explain some interdisciplinary researches and build formal models of intertheoretical relations between their theories in question. In the last stage, analysis of the types of intertheoretical relations and how they work will be a part of Chapter 8.

1.5. Research Plan

This last section systematizes the plan of discussion as follows: Chapter 2 will describe the structuralist theory of science as the basic metatheory of this project and its application. Chapter 3 will deliver some structuralist models of several theories needed from psychology, neuroscience, and artificial neural network. These theories are the Festinger theory of cognitive dissonance, the McCulloch-Pitts neuron, the Hopfield network, the Rosenblatt perceptron, the architecture of artificial neural network, and the delta rule. Chapter 4 will discuss several preparations for the modeling of the intertheoretical connections. They include a revision of the definition of specialization, the notion of an echelon set, and a brief overview of cognitive science. Chapter 5 will deliver a structuralist model of the intertheoretical relations between a specialization of the theory of cognitive dissonance, i.e., for the case of forced compliance dissonance, and the corresponding computational neuroscientific theory. In this modeling, the case of the *dorsal Anterior Cingulate Cortex* investigated by van Veen, et. al. will be used. Chapter 6 will show how the McCulloch-Pitts neuron in neuroscience is in synchronic relation to the Rosenblatt perceptron in artificial intelligence. It will be a model of intertheoretical relation between the models of neurons in both disciplines. Chapter 6 will also deliver a structuralist model of the intertheoretical relations between the Festinger theory of cognitive

dissonance and the Hopfield network according to the consonance model. Chapter 7 will be about building a structuralist model of intertheoretical relations between the theory of forced compliance dissonance and the feed-forward neural network according to the connectionist model. Chapter 8 will discuss the relevance and the contributions of this research for the current state of the related discussion about intertheoretical relations (or respectively reduction) and the unity of science in the philosophy of science and cognitive science. Chapter 9 contains a critical reflection about how this research contributes to the scientific practice in interdisciplinary fields and suggests the possible developments and improvements of this project for some future works.

Chapter 2
The Structuralist Theory of Science and Intertheoretical Connections

This chapter will describe the basic idea of the structuralist theory of science and how it will work to give an account of intertheoretical connections. This topic will be discussed in several parts as follows. The first part will explain the basic concept of formal modeling of scientific theories. The second part will discuss the notion of intertheoretical connection and its kinds. The third part will discuss the results of the intertheoretical links, namely the notions of a theory-net and a theory-holon. Scientific theories are connected to other scientific theories, and this results in two kinds of relation networks. The first kind is a local relation network, which is called a "theory-net." And the second kind is a global one, which is called a "theory-holon." The fourth part of this chapter will discuss the idea of theoreticity with respect to intertheoretical connections. And finally, the idea of a fragment will be discussed in the last part.

2.1. The Basic Concept for Formal Modeling of Scientific Theories according to the Structuralist Theory of Science

The structuralist theory of science was developed by Joseph Sneed, Wolfgang Stegmuller, Wolfgang Balzer, and C. Ulises Moulines. Although many books and articles have been written about this theory, its central book is *An Architectonic for Science*, written by Balzer, Moulines, and Sneed in 1987. This dissertation will use the abbreviation 'BMS' to refer to this book. The structuralist theory of science as a kind of metatheory has scientific theories as its objects of investigation. Therefore, the structuralist theory of science does not (want to) create scientific theories – it assumes that they

already exist. The goal of this metatheory is the modeling and analyzing the deep logical structure of scientific theories. For this goal, the structuralist program applies several formal tools, including, most importantly, set theory and model theory.

The structuralist metatheory of science understands scientific theories as models to explain phenomena. As a model, a scientific theory unifies various aspects of phenomena to explain those phenomena. Under the model concept, the structuralist theory of science, as Bartelborth (1996) depicts it, makes clear in general terms that a model is merely a representation that is specifically designed to serve as a representation of something in the theoretical level of knowledge. A model does not necessarily have to be an isomorphic image of objects and their properties and relations. A model must only be able to show the isomorphic correlations between certain aspects of reality and certain parts of the model (Bartelborth, 1996, p. 364). In the case of a scientific theory, the structuralist theory of science links this intuitive idea of a model mainly to the formal notion of model in logic and mathematics. According to the structuralist approach, logic, and mathematics, especially set theory, represent the terms (or concepts) of scientific theory and the relations between those terms (or concepts). Therefore, the structuralist theory of science under the concept of 'model' always conceives it as a formal or logical-semantic model.

The first and smallest model-theoretic concept for a scientific theory is called 'theory-element' in the structuralist theory of science. The term 'theory-element' is understood as "the smallest unit of empirical science that has all the features required to say something interesting about the world." (BMS, p. xx). Each theory-element contains a vocabulary or conceptual structure and an empirical law-statement or a law-like statement formulated with this vocabulary, and "a specification of the things to which this law is intended to apply" (BMS, p. xx). This theory-element is a construct in terms of set theory and model theory. The basic intuition of this approach is "that

the smallest significant or interesting parts of empirical science are best characterized, not as linguistic entities, but as model-theoretic entities" (BMS, p. xxi).

For building a theory-element, the structuralist theory of science sees that the simplest structure of a scientific theory can be modeled formally as a structure of the form: $\langle D_1, \ldots D_m, R_1, \ldots, R_n \rangle$, where D_i is the basic set and R_j is the relation between these basic sets. The D_i contains the objects of a theory – they can be both empirical objects and purely mathematical objects. The R_j represents the relation or function between the objects of the modeled theory; in the quantitative disciplines, the R_j is often a function of empirical objects to real numbers or vectors. This way of representation serves as the first step for the modeling of scientific theories. The selected sets of axioms determine most precisely those classes of models that represent the specific areas of the phenomena that are important for the theory.

The next step consists in distinguishing among the selected axioms two classes, namely, the class of the 'frame conditions' (which defines the class of the potential models ($\mathbf{M_p}$)) and the class of the substantial laws (which defines the class of the actual models (**M**)). While the potential models contain only the basic concepts and the formal characteristics of the theory – it does not say anything interesting about the world – the actual models, which contain the essential laws of the theory, say something interesting about the world. For the formation of these two kinds of models, the structuralist theory of science uses the method introduced by Patrick Suppes known as 'axiomatization by means of a set-theoretical predicate' (Moulines, 1996, p. 6). The formal definitions of the potential models and the actual models are given as DI-8 and DI-8* in BMS, p.17. From the definition of $\mathbf{M_p(T)}$ and **M(T)**, we can see that the relation between $\mathbf{M_p(T)}$ and **M(T)** is as follows: **M(T)** is a subset of $\mathbf{M_p(T)}$ or $\mathbf{M(T) \subseteq M_p(T)}$. This step characterizes the identity of the scientific theory as the pair $\langle \mathbf{M_p, M} \rangle$. This pair can be called

the ‘model element’ because it provides the essential unit to comprehend the essence of the theory.

However, for some analyses of scientific theories, especially for analysis of their intertheoretical connections, these two classes are not enough. For this purpose, the structuralist theory of science provides us with a fundamental concept of intertheoretical connections (links) – symbolized by ‘**L**’ for such purposes. The structuralist sees that scientific theories are not isolated units, but they are joined together with one another in specific connections or relations. The definition of the links is given in BMS, p. 61.

Another useful tool that the structuralist theory of science provides is the distinction between two different conceptual and methodological levels of a theory, namely the **T**-theoretical level and the **T**-non-theoretical level. The **T**-theoretical level is to be understood as specific concepts of a theory which can be obtained only on the assumption of the theory itself (see the criterion of **T**-theoreticity in BMS, p. 55). Other concepts of the theory, which can be obtained without presupposing the theory itself, belong to the T-non theoretical level. The **T**-non-theoretical concepts come from observations or from other theories. This distinction leads to a new substructure that contains only the non-theoretical elements, the “class of partial potential models,” and is symbolized by $\mathbf{M_{pp}}$ – while the potential models $\mathbf{M_p}$ contain both the **T**-theoretical and the **T**-non-theoretical elements. The class of partial potential models can be obtained by the function $\mathbf{r: M_p(T) \rightarrow M_{pp}(T)}$. The definition of the class of partial potential models $\mathbf{M_{pp}}$ is given in BMS, p. 57.

The potential models $\mathbf{M_p}$, the actual models **M**, the partial potential models $\mathbf{M_{pp}}$, and the global links **GL**, **–** together with the global constraint **GC** and the approximations **A**, which are not discussed here because they are not relevant for this dissertation, – form the core of the theory-element **K**. The definition of a theory-core is given in BMS, p.79. The components of the theory core **K** do not yet reflect the empirical side of empirical theory. For this reason, another domain must be added so that the concept of a theory-

element can represent a real empirical theory. A scientific theory must in principle be applicable to real phenomena so that explanations, predictions, and applications are possible in the form of technologies. This domain is the domain of intended applications and is symbolized by **I**. The characterization of a theory-element is given in the structuralist theory of science by a tuple ⟨**K, I**⟩.

The structuralist theory of science holds three assumptions for the determination of the domain of intended applications. According to the first assumption, the intended applications are neither pure reality nor pure experience. The domain of intended applications does not contain pre-conceptual things or sense-data. According to this, the domain of intended applications of a theory is conceptually determined by the existing terms. According to the second assumption, the intended applications of the theory are not concerned with all human experience, but with local and diverse parts of human experiences. The structuralist concept of science does not assume there exists a kind of theory of everything. Each scientific theory has its domain of intended applications **I**; these domains of different theories can overlap, partially overlap, loosely connect, or be completely apart. The structuralist theory of science regards the domain of intended applications as a subclass of the partial potential models ($\mathbf{M_{pp}}$). This approach is a weak characterization of the intended applications. To discuss the domain of intended applications in more detail, it cannot be stated in the pure formulation utilizing set theory or model theory, since the intended applications are not independent of historical and pragmatic factors. The definition of the intended application is given in BMS, p. 88.

Since scientific theories are not isolated units but are related to other theories, the theory-elements, as their formal models, appear in groups and are connected through intertheoretical connections (links). Intertheoretical links serve to transmit ‘information’ between theory-elements (BMS, p. xx).

Certain intertheoretical connections such as specializations link theory-elements with the same vocabulary, and different laws, in networks that represent another general conception of a scientific theory. Moreover, global empirical science is a network of (all) theory-elements that connect with other theory-elements through various intertheoretical connections **L**. In this way, the structuralist theory of science divides the formal explication of the intuitive idea of a scientific theory into three different concepts. The simplest level is the idea of a theory-element. In the next level, some (at least two) theories having the same basic conceptual apparatus appear connected and form a theory-net. After that, many (though not all) theories in the second sense (i.e., theory-nets) form a global structure of scientific theories called a 'theory-holon.'

In real scientific practice, scientific theories can change over time in three directions: "First, the things to which the laws in individual theory-elements are expected to apply ... may grow or diminish. Second, theory-elements may appear and disappear from the complex. Finally, intertheoretical links between the theory-elements may appear and disappear from the complex" (BMS, pp. xx–xxi). In these transformations of scientific theories and intertheoretical connections already discussed, two kinds of representations of the links between scientific theories are considered. First, there is the concept of 'theory-evolution.' A theory-evolution is a development of 'theory-nets' through time. Secondly, the synchronic representation shows how many scientific theories can be assembled at a time and can be linked to each other to provide explanations of certain phenomena or specific intended applications. This dissertation focuses on this synchronic representation of intertheoretical relations.

2.2 The Concept of Intertheoretical Connections and Their Varieties

2.2.1. The Concept of Intertheoretical Connections

Scientific theories are always connected to other scientific theories. Therefore, the structuralists understand intertheoretical connections in two ways, i.e., intertheoretical connections as a kind of bridge connecting several scientific theories and intertheoretical connections as essential components of each scientific theory.

As one of the essential components of a scientific theory, the structuralist theory of science includes the concept of intertheoretical connection in the concept of theory-core. Intertheoretical connections belong to the concept of a scientific theory because the idea of an isolated theory in science is fundamentally deficient. A scientific theory's identity can only be adequately understood if one considers its links with other scientific theories.

As intertheoretical bridges, intertheoretical relations are not relations between statements or sets of statements but relations between models or sets of models. In this case, it is important to formulate more precisely, in a formal way, either that a connection is a bridge between models of the same theory or a bridge between models of different theories. The first is called intratheoretical bridge or constraint (**C**), while the second is an intertheoretical bridge or link (**L**). The definition of a bridge can be seen in Moulines & Polanski, 1996, p. 220.

As stated above, a model element $\mathbf{E} = \langle \mathbf{M_p^i}, \mathbf{M^i} \rangle$ is the smallest pair that forms the identity of a scientific theory. If there is an intertheoretical connection, such as $\boldsymbol{\lambda} \subseteq \mathbf{M_p^{i'}}\ \mathbf{M_p^j}$, it is convenient to write the intertheoretical connection between two different theories as $\mathbf{E^i \lambda E^j}$. If there are two models to be considered, x^i and x^j, the intertheoretical connection between them is $\langle x^i, x^j \rangle \in \boldsymbol{\lambda}$, and can be written as $x^i \boldsymbol{\lambda} x^j$.

2.2.2. The Varieties of Intertheoretical Connections

Because it is important to understand the nature of intertheoretical relations between scientific theories, the structuralist theory of science provides us with many tools for the analysis of intertheoretical relations. The structuralist metatheory of science identifies various kinds of intertheoretical connections as relations between scientific theories. According to Moulines (1992), two basic intertheoretical connections are the entailment intertheoretical connection and the determining intertheoretical connection. All other intertheoretical connections can be derived from these two types either by adding further conditions or by combining them. There are at least five types of specific intertheoretical connections formally characterized by the structuralist metatheory of science for synchronic intertheoretical relations. Some of them can be attributed to the entailment intertheoretical connections because they are specific types of entailment links – specialization, reduction, equivalence, and approximation. However, the partial reduction is a specific type of determining link.

In addition to these five specific synchronic intertheoretical connections, Moulines (2014) also characterized four specific diachronic intertheoretical connections that capture the dynamic development of a scientific theory, namely crystallization of a theory, evolution of a theory, embedding of one theory into another one, and replacement of one theory by another accompanied by partial (semantic) incommensurability. Since diachronic intertheoretical connections are not the topic of this dissertation, they will not be discussed here. This notwithstanding, the case discussed in Chapter 6 is historically a case of intertheoretical embedding, namely the embedding of the McCulloch-Pitts neuron into the Rosenblatt perceptron. This intertheoretical connection is a connection from the model of a neuron in neuroscience into the mathematical model of a neuron, called perceptron, that is used in the artificial neural network for the simulation of cognitive dissonance.

2.2.2.1. Entailment Intertheoretical Connections (Links)

Entailment intertheoretical connections are intertheoretical connections that connect the whole of the actual models of two theories. Therefore, they form a unity of classes from the structures of different theories. The formal definition of entailment intertheoretical connections is to be found in Moulines & Polanski, 1996, p. 223. This definition of the entailment intertheoretical connections (links) can also be applied to the analysis of theory-holons by adding some additional conditions for the **T**-non-theoretical level or an intended application (See DVIII-2 in BMS, p. 392). The additional conditions are as follows: (1) the entailment link must connect the actual models of both theories, and (2) the entailment link provides a mapping to the set of intended applications of one of both theory-elements according to the function r. This mapping defines the local empirical claims of the links in the theory-holon.

2.2.2.2. Determining Intertheoretical Connections (Links)

Determining intertheoretical connections represent relations between single terms or concepts (i.e., they are term to term relations), which connected theories contain. The formal definition of determining intertheoretical connections is given in Moulines & Polanski, 1996, p. 223.

2.2.2.3. Intertheoretical Specialization

Specialization is another type of intertheoretical connections that arise because of an addition of special laws or law-like statements to a theory **T**, such that a new theory **T'** comes into existence. The additional law(s) cause an improvement of the explanatory power with the cost of a limitation of the explanatory range. From the model-theoretical point of view, this addition of special law(s) into the existing laws can be considered as creating a subset **M'(T)** from the current models **M(T).** The subset **M'(T)** satisfies more constraining conditions for a partial set of **I(T)** in a more limited empirical

range. This notion of specialization is defined formally as DIV-1 of BMS, p. 170.

A specialization relation has the following characteristics: (a) the specialization is used for special cases, (b) the specialization limits the empirical content of the initial theory-element core. These properties are expressed in BMS, p. 170.

2.2.2.4. Intertheoretical Theoretization

The essential idea of theoretization can be understood through the difference between **T**-theoreticity and **T**-non-theoreticity. The **T**-theoretical concepts of a theory are the concepts that can only be determined by presupposing the theory itself. The other concepts, which do not require the theory **T** itself for their determination, are called **T**-non-theoretical concepts. The **T**-non-theoretical concepts can be derived from other theories, measurements, and observations. Here we can easily see that a **T***-non-theoretical concept of theory **T*** can be a **T**-theoretical concept of the theory **T** if the theory **T*** is linked to theory **T** or applies or accepts certain concepts of theory **T**. Thus, whether a concept is theoretical or non-theoretical, is not an inherent property of the concept. The identity of a concept can change from one theory to another through such intertheoretical relations. Such an intertheoretical relation is called theoretization. More precisely, the concept of theoretization can be understood as follows: **T*** is a theoretization of **T**, if **T***-non-theoretical concepts are concepts of **T**, either **T**-theoretical or **T**-non-theoretical. There is a distinction between two cases: **T*** is a theoretization of **T** in the weak sense, if some of the **T***-non-theoretical concepts come from **T**; while **T*** is a theoretization of **T** in the strong sense, when all **T***-non-theoretical concepts come from **T**. The intertheoretical theoretization is defined formally as DVI-1 in BMS, p. 251.

2.2.2.5. Intertheoretical Reduction

The intertheoretical reduction is one of the most important topics in many discussions of the philosophy of science. There are several types of reductions in these discussions. The first type is the historical reduction, which can be comprehended intuitively by considering historical developments of specific scientific explanations, such as the development of the explanation of the motion of planets from Ptolemy via Copernicus and Kepler to Newton. Historically, these relations were sometimes shown in the following circumstances: a theory **T** is replaced by a new and conceptually different theory **T*** with related or similar, but better constructed intended applications (BMS, p. 252). There are intense discussions about the justification of this transition, but this discussion is no longer the subject of this work because this work is not intended to discuss the diachronic intertheoretical relations. The second kind of reduction that can be made is due to the simplification of applications of theory. Such a reduction is called 'practical reduction.' In many interdisciplinary cases, there are also reductions due to the speculative view that one particular area is viewed as more basic than another. However, this speculative mode of reduction will not be discussed in this work, although interdisciplinarity is at the center of my research. The distinction between historical reduction, practical reduction, and speculative reduction plays no role in our discussion about reduction as long as the modeling is concerned. What is crucial is the distinction between exact reduction and approximative reduction. The expression "reduction" used here denotes the exact reduction, and the expression "approximation" refers to the approximate reduction, which can be modeled by implementing the notion of intertheoretical approximation grounded on exact reduction. For simplicity and clarity of modeling intertheoretical connections, I will not consider the notion of approximation in the case studies in this dissertation.

The reduction discussed here is a kind of intertheoretical relation between the structural classes of the theories and not simply between the

concepts or the terms of the theories. For a formal definition of the reduction relation between a **T** 'the reduced theory' and a **T*** 'the reducing theory,' the structuralist metatheory of science has considered seven conditions for an adequate reduction relation. (See BMS, pp. 275–276) However, we do not discuss these details here, since they are not relevant to the present discussion.

2.2.2.6. Intertheoretical Equivalence

In discussing the fourth kind of intertheoretical connections, the so-called equivalence, the structuralist distinguishes two kinds. The first is empirical equivalence, while the second is theoretical equivalence or simply equivalence.

Empirical Equivalence. Empirical equivalence focuses on the T-non-theoretical level of the theory, namely the intended applications. This approach considers theories as a tool for explaining particular phenomena that they take up or for solving problems in the area of the phenomena. In this approach, the structures of the complete theoretical instruments can be neglected as long as both theories provide the 'same' explanations and solve the 'same' problems with the 'same' systems or if they have the 'same' empirical content. This empirical equivalence is formally defined in BMS, on pages 288-289 as D VI-9 and TVI-7. By giving the formal definition, the structuralist theory of science assumes that empirical equivalence is a kind of global intertheoretical relations without further investigation of its components.

Theoretical Equivalence. Theoretical equivalence also takes the theoretical concepts of the theories into account, which belong to the potential models. Two theories are equivalent if their complete theoretical structures are in some respects isomorphic, and the two theories explain the 'same' phenomenon. Moreover, this relationship between potential models provides a satisfactory comparison between the intended applications. The relation,

however, does not have to be bijective between models. This equivalence relation is formally defined as DVI-11 in BMS, p. 297.

2.2.2.7. Intertheoretical Approximation

The last special type of the intertheoretical connections characterized in BMS is intertheoretical approximation. The term "intertheoretical approximation" is understood as a kind of inexact intertheoretical relation between two theory elements **T** and **T'**, where $\mathbf{M_{pp}(T) \neq M_{pp}(T')}$ and $\mathbf{M_{p}(T) \neq M_{p}(T')}$. The two theoretical elements, which are compared or connected here by the intertheoretical approximation, belong to two synchronically different theories. Therefore, the intertheoretical approximation is a relation between theory elements in the form $\langle \mathbf{K, A, I} \rangle$ and $\langle \mathbf{K', A', I'} \rangle$, where **K** and **K'** are conceptually different, and **A** and **A'**, which are the sets of admissible blurs of both theories **T** and **T'** related to the intertheoretical connection between them, are also different. If $\mathbf{M_p}$ and $\mathbf{M_p'}$ are different, then **A** and **A'** will also be different. Therefore, there are no "common" blurs that express the relation between the two theories.

2.3. Modeling Relations (or Networks) between Theories

After discussing various kinds of intertheoretical connections, now we can move on to discussing how extensive intertheoretical relations can be produced from these various kinds of intertheoretical connections. The structuralist metatheory of science has already characterized two products of intertheoretical relations, namely: theory-nets and theory-holons.

2.3.1. Theory-nets

With the term 'theory-net,' the structuralist metatheory of science describes a relation between two or more theory-elements with the same potential models and same partial potential models and are related through a

particular intertheoretical connection, i.e., specialization(s). The notion of a theory-net corresponds to a 'local' idea, that is, the combination of a scientific theory with other closely linked theories. A standard example is classical particle mechanics with its specializations such as Newtonian classical particle mechanics, Hooke's classical particle mechanics, and others. The notion of theory-net is defined in DIV-2 of the BMS, p. 172. The specialization connection in theory-nets is a partial order relation; that is, it has the following properties: reflexive, transitive, and antisymmetric (BMS, p.172).

There are various types and cases of nets of scientific theories – we can represent them by graph theory. In the normal situation, a theory-net (**N**) consists of at least two different theory-elements, which are connected to each other. The cases must be that they are either specializations of another common 'higher' theory-element or else one is a specialization of the other. In such a theory-net, there is always at least one theory-element that is not a specialization of other theories, and there are also theory-element(s) that have no specializations. Theory-elements, which are not a specialization of other theory elements, are called the basic theory-element(s). An important type of theory-net is the theory-net with a single basic theory-element. In graph-theoretical representation, it forms a tree-like structure. Therefore, it is called a theory-tree. This kind of theory-net, the basic theory that forms it, and the condition of connectedness are discussed and defined in BMS, pp. 173–175.

The structuralist metatheory of science conceives a theory element as a pair ⟨**K**, **I**⟩. Therefore, the relations between the cores and the intended applications of the theories involved in the network should also be considered in the construction of a theory-net. The relations between the cores and the relations between the intended applications of the theories in the network are formulated as DIV-6 and D IV-7 in BMS, pp. 176–177. In a tree-like network, the net of cores and the nets of applications have the same net structure as the

original theory-net, if there is only a set of the intended applications for each core in the original network, and vice versa (BMS, p. 177 see T IV-5).

Since the empirical content of a theory-element is interpreted as the statement $\mathbf{I} \in \mathbf{Cn(K)}$, the theory-net has as many individual empirical contents as the theory-elements in the net. BMS summarizes the global empirical content of the net as the conjunction of all these individual empirical contents of the theory-elements, as formulated in D IV-8 (BMS, p. 177).

In the case of an unconnected theory-net with possibly unrelated sub-nets, its empirical claim contains more or less amorphous conjunction of individual empirical claims. However, in the case of the connected theory-net, the empirical claim is important to be considered inasmuch as all individual claims refer to the same $\mathbf{M_{pp}}$. All the individual sets of intended applications $\mathbf{I_i}$ of theory-elements in such a connected, tree-like net, are the subsets of the basic set $\mathbf{I_0}$. These individual subsets $\mathbf{I_i}$ can be subsumed under the basic core $\mathbf{K_0}$ as supplemented by the addition of certain restricting conditions to the basic core $\mathbf{K_0}$. These specific conditions are the conditions that define $\mathbf{K_i}$. However, we have to consider that the empirical claim of the basic theory-element can be vacuous because of $\mathbf{Cn(K_0)} = \text{Po}\ (\mathbf{M_{pp}})$. However, even in such a case, the global empirical assertion of the network may not be vacuous because of $\mathbf{Cn(K_i)} \subset\neq \text{Po}\ (\mathbf{M_{pp}})$ at least for some specializations $\mathbf{K_i}$. At any rate, even if the basic statement is not vacuous, it usually is very weak.

2.3.2. Theory-holons

The relations between some theories of different theory-nets, which enter further intertheoretical connections, build a 'theory-holon.' The concept of theory-holon contains a 'global' idea. In a theory-holon, scientific theories have connections not only with their close relatives, but also with many other theories from other areas of empirical science, be it within one and the same discipline, or else from different disciplines. The connections here are much

more complicated than in a theory-net because many different intertheoretical connections are involved. There are even several types of intertheoretical connections between two theories **T** and **T'** in such a global relation. To make it simple, the discussion is firstly held as if there is always only one connection. This approach is based on reality that the conjunction of several links – set-theoretically expressed by their intersection – is also a link. The general idea of theory-holon is defined as D VIII-1 in BMS, p. 389. The notation **N** used in the original definition in BMS is replaced by **H** in this dissertation to differentiate the theory-holons from the theory-nets.

In a theory-holon, the intertheoretical connections λ are a partial function mapping a theory-element **T** to another theory-element **T'**. As a partial function, λ implies that there is at most one link between **T** and **T'** and there is room for pairs of theory-elements in **H** that are not connected. The pair $\langle \mathbf{T}, \mathbf{T'} \rangle$, as a domain of λ, means that λ is a subset of the relation between the potential models of both theories, namely $\mathbf{M_p(T)}$ and $\mathbf{M_p(T')}$. Defining a theory-holon requires that all theory-elements in **H** must be connected at least to a theory-element in **H**. Any theory-element which does not satisfy this requirement is called "isolated" and has no connection with a holon. Therefore, there is the possibility of linking many theory-elements in networks through different links because of the transitivity of the intertheoretical connections – if a theory-element **T** is connected to another theory-element **T'** by λ and **T'** is connected to **T"** by λ', then we can always define a new λ" as link between **T** and **T"**.

The network structure of the theory-holon is more complicated than the structure of theory-nets. It contains the global way of connecting various theory-elements. Graph theory represents the structure of the relations created by links. This network of theories $\langle \mathbf{H}, \leq \rangle$ consists of binary relations between theory-elements and can also be expressed as follows: $\mathbf{T} \leq \mathbf{T'}$ iff $\langle \mathbf{T}, \mathbf{T'} \rangle \in$ Dom

(λ). In BMS, pp. 393–394, we can find the definition of a theory-holon in the graph-theoretic expression.

Although a theory-holon deals with global intertheoretical relations, it does not refer to the global empirical claim of the holon, nor to its (global) intended applications. The problem with the global notion of empirical claims and intended applications is that on the practical level of such notions, they are implausible and impractical. In practice, it is more plausible if we refer only to the empirical claim(s) and the intended application(s) of a small piece of the holon. For this reason, we discuss the local empirical claims and the intended applications of a theory-element or a theory-net in the context of the theory-holon, including them.

To discuss the local empirical claims and the 'local' intended applications, we must begin with an (individual) theory-element. Then, the examination of the theory-elements that contribute to the theory-holon follows. In this context, an interpreting link must be considered. By "interpreting link," we mean an intertheoretical connection, which gives interpretations about the non-theoretical concepts of a theory-element. Thus, a theory-element **T'** interprets another theory-element **T** (in holon **H**) iff. ⟨**T'**, **T**⟩∈Dom(λ), and λ(**T'**, **T**) is an interpretive link. In the holon, a theory-element may play both an interpreting and an interpreted role (BMS, p. 396).

2.4. The Intertheoretical Connections and the Concept of T-Theoreticity

As a result of the discussion about the notion of a theory-core, a criterion about the difference between **T**-theoreticity and **T**-non-theoreticity must be present in order to identify the theoretical level conceived as the potential models and the practical or non-theoretical level conceived as the partial potential models. In contrast to the classical metatheory of science, the structuralist metatheory of science assumes that not only the observational concepts belong to the practical level, but also the concepts adopted from

other theories. Therefore, the criterion of T-theoreticity can be sharpened through the relations of theory with other theories in the same holon.

2.5. The Fragment

Although the concept of a fragment does not belong to the theoretical part of the structuralist metatheory of science, BMS discusses this for practical reasons. The term ‘fragment of empirical science’ refers to a part of empirical science that serves as a unit of empirical science in the usual discussions of science. A fragment consists of a few theory-elements connected to each other in specific networks or in a particular scientific field. A fragment is only a part of a theory-holon, but it is larger than single theory-elements or theory-nets. An example is what I will analyze in this dissertation: The fragment of psychology, including the theory of cognitive dissonance as one of its parts, is connected not only with its specializations in the theory-net but also with other theories in psychology. Moreover, as an interdisciplinary field in science, cognitive science forms a larger fragment that connects fragments of psychology with fragments of neuroscience and with fragments of artificial intelligence. For such a practical reason, and for delimiting the discussion, the idea of a fragment of empirical science will be needed later.

Chapter 3
Structuralist Models of Several Scientific Theories in Cognitive Science: The Case of Dissonance Reduction in the Cognitive Process

This chapter will discuss the building of set-theoretical models of several theories in various scientific fields in cognitive science. Since cognitive science is a vast discipline, it will be realistic if this work limits itself only to the relations between several theories from two or three connected fields for explaining a specific case of phenomena. Thus, this dissertation will focus on the phenomena of dissonance reduction in the cognitive process. These phenomena are explained well by the theory of cognitive dissonance from Leon Festinger. There are also many kinds of research in neuroscience and simulation by using artificial neural networks related to these phenomena. This chapter will present several structuralist models of several theories from these fields before we model the intertheoretical connections for such research.

3.1. Psychology

The first scientific field of cognitive science that will be modeled here is psychology, especially the theory of cognitive dissonance from Leon Festinger in 1957. Rainer Westermann has already built the structuralist models for the Festinger theory of cognitive dissonance and its specializations in 1989 and 2000. The content of the theory will be described first and then followed by the structuralist models built by Rainer Westermann.

3.1.1. A Brief Description of the Theory of Cognitive Dissonance

Leon Festinger develops the theory of cognitive dissonance in his book *A Theory of Cognitive Dissonance*, first published in 1957, and then republished in 1985. The basic idea of the theory is that we have an inner drive to hold all attitudes and beliefs in harmony, and therefore, we have tendencies to avoid, reduce, or eliminate disharmony or dissonance (McLeod, 2008, updated 2014). This dissonance theory has great importance and influences in the psychology of motivation and social psychology and has led to many experiments and considerable theoretical progress. The book itself consists of 10 chapters. Chapter 1 presents the general theory of cognitive dissonance, whereas the rest present the specific modifications of the theory in four different specific situations, namely (1) Post-decision Dissonance (Chapters 2 and 3), (2) Forced Compliance Dissonance (Chapters 4 and 5), (3) Dissonance and Information Exposure (Chapters 6 and 7), and (4) Social Disagreement Dissonance (Chapters 8–10). This structure fits well the structuralist idea of theory-element and theory-nets, which are connected by the kind of intertheoretical connection called specialization. This dissertation will only use the theory of cognitive dissonance itself and the forced compliance dissonance for modeling and analyzing the intertheoretical connections.

The Main Theory of Cognitive Dissonance. The theory of cognitive dissonance begins with the idea that "the individual strives toward consistency within himself" (Festinger, 1985, p.1). However, in real life, there are many occasions where a person makes decisions or behaves in ways that cause inconsistencies with her other existing knowledge, opinions, or beliefs about the environment, about herself, or about her behavior, for example, smoking and the knowledge of its side effects. For his theory, Festinger replaces the word 'consistency' by 'consonance' and the word 'inconsistency' by 'dissonance' (Festinger, 1985, pp. 2–3). The hypothesis of the theory is as follows: (1) the existence of dissonance will motivate a person to reduce the

dissonance in order to achieve consonance and (2) when dissonance is present, the person will actively avoid situations and information which would likely increase the dissonance.

The dissonance may occur in the two following situations: (1) new events or new information becomes known to a person that does not have complete control over them. These events or information creates a momentary dissonance with the existing knowledge, opinion, or cognition concerning her behavior or decision. (2) the dissonance can also arise in an everyday situation, where only a few things are clear-cut enough related to behavior or decision about right or wrong, safe or dangerous, etc. There are widely various situations in which dissonance is nearly unavoidable (Festinger, 1985, p. 5). To reduce, or respectively to eliminate the dissonance, there are two possible options for the person. "He might simply change his cognition about his behavior by changing his action, … [or] he might change his 'knowledge' about the effect of [his actions]" (Festinger, 1985, p. 6).

According to Festinger, there are three possible kinds of relations between two cognitions: irrelevance, consonance, and dissonance. Two cognitions are irrelevant if they have nothing to do with each other. "Under such circumstances where one cognitive element implies nothing at all concerning some other element, these two elements are irrelevant to one another" (Festinger, 1985, p. 11). Two cognitions are in dissonance if they are inconsistent or contradictory to each other according to cultural or specific group standards. Otherwise, they are in consonance (Festinger, 1985, p. 13). The reasons for dissonance could be a logical inconsistency, cultural custom and manners, a specific opinion about particular more general opinion(s), experience(s) in the past, etc. (Festinger, 1985, p. 14).

"All dissonance relations are not of equal magnitude, [therefore] it is necessary to distinguish the degree of dissonance and to specify what determines how strong a given dissonance relation is" (Festinger, 1985, p. 16). The notion of the magnitude of dissonance is defined by the importance of

the elements of cognition. “If two elements are dissonant with one another, the magnitude of the dissonance will be a function of the importance of the elements. The more these elements are important to, or valued by, the person, the greater will be the magnitude of a dissonance relation between them” (Festinger, 1985, p. 16).

This magnitude of dissonance is an essential variable in determining the pressure to reduce dissonance. “The strength of the pressures to reduce the dissonance is a function of the magnitude of the dissonance” (Festinger, 1985, pp. 17–18). In his book, Festinger mentions several options for a person to reduce dissonance, namely: (1) by changing a behavioral, cognitive element, (2) by changing an environmental, cognitive element, or (3) by adding new cognitive elements. Despite these possible ways of reducing dissonance, the attempts to reduce or eliminate dissonance are not always successful. Some dissonance might have resistance against these attempts. This success or failure to reduce dissonance defines the maximum magnitude of the dissonance. “The maximum dissonance that can exist between any two elements is equal to the total resistance to change of the less resistant element. The magnitude of dissonance cannot exceed this amount because, at this point of maximum possible dissonance, the less resistant element would change, thus eliminating the dissonance” (Festinger, 1985, p. 28). Besides attempting to reduce or eliminate dissonance, a person also has tendencies to avoid the increase of dissonance.

The Theory of Forced Compliance Dissonance. Sometimes a person behaves in a manner counter to her convictions or will publicly make a statement, which she does not really believe. She does it because of public compliance – in the form of threat of punishment or special rewards – without accompanying changes of private opinion (Festinger, 1985, p. 85). It will increase dissonance and the pressure to reduce it.

For recognizing a case of public compliance in changes of a person’s private opinion from her genuine opinion, there are two strategies according

to Festinger, namely: (1) by removing the source of influence or pressure, and (2) by direct measurement of her private opinion that can be done by assuring the person's anonymity – by saying that her identity will not be exposed.

The strength of dissonance is determined by the number and importance of cognitive elements that are dissonant with the cognition about the overt behavior. It is also determined by elements of cognition that "correspond to the knowledge that a reward has been obtained or that a punishment has been avoided" (Festinger, 1985, p. 90). Both determining elements must be in one of the following relations: (1) "The expected reward or punishment had to be sufficient, in relation to the resistance to change, to produce the compliant behavior in the first place. Consequently, it is reasonable inference to suppose that the sum of consonant relations is greater than the sum of dissonance. ... However, if the reward or the punishment is a too great reward or punishment, dissonance will be small" (Festinger, 1985, p. 91). The smaller the rewards or punishment is, the higher the magnitude of dissonance is. (2) Alternatively, the expected reward or punishment is too small to produce overt behavior so that the person stays with his private opinion or decision. In this case, the bigger the reward or punishment is, the higher the dissonance that will be produced. (Festinger, 1985, pp. 91–92)

Festinger suggests some strategies to reduce this specific kind of dissonance as follows: (1) by adding the weight of reward and punishment, or (2) by specific actions to change the private opinion or its – as much as possible – cognitive elements becoming consonant to the overt action.

3.1.2. A Structuralist Model of the Theory of Cognitive Dissonance

For a structuralist model of Festinger's theory of cognitive dissonance, this dissertation just takes over the models, which are built by Rainer Westermann. These models can be found in two of his works, namely *Festinger's Theory of Cognitive Dissonance: A Revised Structural Reconstruction* (1989) and *Festinger's Theory of Cognitive Dissonance: A*

Structuralist Theory-Net (2000). The first work focuses more on the building of the model and the considerations behind it, whereas the second focuses more on the model and its specializations.

In the first work, Westermann builds two models for Festinger's theory of cognitive dissonance, namely: the full version, called DissA, and the simplified version, called DissB. This research uses the simplified version for two reasons: (1) according to Westermann DissA does not adequately represent the theoretical basis of all empirical studies on dissonance theory for two reasons: (a) empirical research refers to additional terms and relationships, and (b) empirical dissonance research does not make use of all terms and relations of element DissA. (2) DissB represents the common theoretical reference point of all dissonance research and is hypothesized to be the basic model of all theory-elements from which various parts and versions of dissonance theory can be reconstructed.

3.1.2.1. The Theory-Element of the Theory of Cognitive Dissonance (DissB)

In building his structuralist model, the DissB, Westermann considers the following concepts as the elements of the potential models ($\mathbf{M_p}$), respectively, the following basic concepts. Firstly, the most fundamental concept of the theory is “cognition” or “cognitive element” (Westermann, 1989, p. 34), which means “any knowledge, opinion, or belief about the environment, about oneself, or about one's behavior” (Festinger, 1985, p. 3). To define the cognition that plays an essential role in increasing or reducing dissonance in a particular moment, Westermann differentiates between the set of relevant cognitions that are present (*Cognition*) and the set of interesting raw elements of cognition of a specific subject or group (*RawCog*). *RawCog* refers to the set of cognitions belonging to the subject or the group. *Cognition* is the set of cognitions that plays a role in dissonance or consonance. This differentiation is made by the time when the dissonance takes place or

increases. Therefore, *Cognition* is defined as a subset of *RawCog*. An auxiliary set for this task is the set of time points (*Time*),

Secondly, the other fundamental concepts are the concepts of possible relations between several cognitions or cognitive elements: consonance, dissonance, and irrelevance. In the construction, according to Westermann, only the two first concepts are relevant. *Conscog* and *Disscog* model the concepts of consonant and dissonance relations between pairs of cognitive elements. "The subsets *Disscog* and *Conscog* of the cartesian product Cognition × Cognition are mutually exclusive and encompass the dissonant and the consonant pairs of cognitions, respectively" (Westermann, 2000, p. 191).

The next important concepts are the degree of dissonance or consonance. These concepts are *pairdiss* for the degree of dissonance and *paircons* for the degree of consonance. These functions map the elements of *Disscog* and *Conscog* into a set of positive real numbers representing the degree of the relations. Westermann adds a note: "these functions are defined so that the degree of dissonance and consonance may vary over time" (Westermann, 2000, p. 191).

The magnitude of dissonance and the magnitude of consonance in Festinger's theory depend on the degree of importance of the relationship. Therefore, Westermann introduces a function *pairimp*. With this function, the degree of importance is represented by a positive real number. The magnitude of dissonance and the pressure to reduce it are represented by the functions *diss* and *redpress* that attribute a positive real number to each element of Cognition for their measure.

Finally, the auxiliary functions *confl* and *suppo* (magnitude of conflict and support) are defined by Westermann as sums of the value of importance. "For each element of *Cognition*, the summation runs overall relationship to other cognitive elements that are dissonant or consonant, respectively. The

support of an element also represents the resistance to change for that cognition" (Westermann, 2000, p. 192).

From the concepts above, Westermann builds the potential model as follows (Westermann, 2000, p. 190):

DIII-1: x is a potential model of Festinger's theory of Cognitive Dissonance ($x \in \mathbf{M_p}(\text{DissB})$) iff there exist Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress such that:

(1) $x = \langle \text{Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress} \rangle$

(2) $\text{Time} \subseteq \text{IR}$ is a finite, non-empty set of points of time

(3) Rawcog is a finite, non-empty set of raw elements of cognition

(4) $\text{Cognition} \subseteq \text{Rawcog} \times \text{time}$ (actual elements of cognition)

(5) $\text{Disscog} \subseteq \text{Cognition} \times \text{Cognition}$ (dissonant cognitions)
$\text{Conscog} \subseteq \text{Cognition} \times \text{Cognition}$ (consonant cognitions)
$\text{Disscog} \cap \text{Conscog} = \varnothing$

(6) $\text{pairdiss: Disscog} \rightarrow \text{IR}_0^+$ (dissonance within pairs)
$\text{paircons: Conscog} \rightarrow \text{IR}_0^+$ (consonance within pairs)

(7) $\text{Pairimp: (Disscog} \cup \text{Conscog)} \rightarrow \text{IR}_0^+$ (importance of pairs)

(8) $\text{diss: Cognition} \rightarrow \text{IR}_0^+$ (magnitude of dissonance)
$\text{redpress: Cognition} \rightarrow \text{IR}_0^+$ (dissonance reduction pressure)

(9) $\text{confl}(c_{it}) := \Sigma_{(cit,\, ckt) \in \text{Disscog}} \text{pairimp}(c_{it}, c_{kt})$ (degree of conflict)

(10) $\text{suppo}(c_{it}) := \Sigma_{(cit,\, ckt) \in \text{Conscog}} \text{pairimp}(c_{it}, c_{kt})$ (degree of support)

The next crucial step in building a structuralist model of the theory of cognitive dissonance is building an actual model, which contains the law-statements or the law-like statements of the theory. Westermann uses the following indexing system to refer to typical elements of *Cognition*: c_{it} with $i \in$ *Rawcog* and $t \in$ *Time.* (1) The first law statement or law-like statement is if two cognitive elements are dissonant with one another. The magnitude of the

dissonance “will be a function of the importance of the elements” (Festinger, 1957, p. 16 in Westermann, 2000, p. 193) and “increases as the importance or value of the elements increase” (Festinger, 1957, p. 18 in Westermann, 2000, p. 193). Westermann follows Festinger’s assumption of “a strictly monotone increasing relationship between importance and dissonance of the pairs of cognition” (Westermann, 2000, p. 193). (2) The second law statement or law-like statement is that the same is also the case for the consonant relation between a pair of cognitive elements.

(3) The magnitude of dissonance between the element in question and the remainder of the person's cognition, “will *depend on* the proportion of relevant elements that are dissonant with the one in question” (Festinger, 1957, p. 17 in Westermann, 2000, p. 194) and “*is a function of* the weighted proportion of all relevant relations … that are dissonant. The term ‘weighted proportion’ is used because each relevant relation would be weighted according to the importance of the elements involved in the relation” (Festinger, 1957, p. 18 in Westermann, 2000, p. 194). The two functions *confl* and *supp* can be used to formulate this relationship. Here, Westermann assumes that ‘depend on’ and ‘is a function of’ have a strictly monotone relationship.

(4) According to Westermann, the central point of Festinger's theory refers to the consequences of dissonance arousal: “The presence of dissonance gives rise to pressures to reduce dissonance. … The strength of the pressure to reduce the dissonance is a function of the magnitude of the existing dissonance” (Festinger, 1957, p. 263 in Westermann, 2000, p. 194). The existing dissonance between cognitive elements can be reduced or eliminated by changing one of these elements, by adding new elements, or by decreasing the importance of the elements involved. The activity of reducing dissonance does not ensure that dissonance will be reduced; sometimes, it can even be increased. The dissonance theory does not predict this, but only says “that in the presence of a dissonance, one will be able to observe the attempts

to reduce it" (Festinger, 1957, p. 24, in Westermann, 2000, p. 194). The function *redpress* represents the magnitude of these *attempts*.

The actual models of the Festinger theory of cognitive dissonance can be defined as follows: (Westermann, 2000, p. 193)

DIII-2: x is an actual model of the Festinger theory of cognitive dissonance ($x \in \mathbf{M}(\text{DissB})$), iff there exist Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress such that:

(1) $\mathbf{x} = \langle$Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress$\rangle \in \mathbf{M_p}(\text{DissB})$

(2) For all (c_{it},c_{jt}), $(c_{ku}, c_{lu}) \in$Disscog: if pairimp $(c_{it},c_{jt}) <$ pairimp(c_{ku},c_{lu}), then pairdiss$(c_{it},c_{jt})<$pairdiss(c_{ku}, c_{lu})

(If the importance of the pair c_{ku} and c_{lu} is greater than the importance of the pair c_{it} and c_{jt}, then the dissonance of the pair c_{ku} and c_{lu} is greater than the dissonance of the pair c_{it} and c_{jt})

(3) For all (c_{it}, c_{jt}), $(c_{ku}, c_{lu}) \in$Conscog: if pairimp $(c_{it},c_{jt}) <$ pairimp(c_{ku},c_{lu}), then paircons$(c_{it},c_{jt})<$paircons(c_{ku}, c_{lu})

(If the importance of the pair c_{ku} and c_{lu} is greater than the importance of the pair c_{it} and c_{jt}, then the consonance of the pair c_{ku} and c_{lu} is greater than the consonance of the pair c_{it} and c_{jt})

(4) For all c_{it}, $c_{ju} \in$Cognition: if confl(c_{it})/(confl(c_{it}) + suppo(c_{it})) < confl(c_{ju})/(confl(c_{ju}) + suppo (c_{ju})), then diss(c_{it}) < diss(c_{ju}).

(If the proportion between the degree of conflict of c_{ju} and the sum of the importance of c_{ju} is greater than the proportion between the degree of conflict of c_{it} and the sum of the importance of c_{it}, then the dissonance of c_{ju} is greater than the dissonance of c_{it})

(5) For all c_{it}, $c_{ju} \in$Cognition: If diss(c_{it}) < diss(c_{ju}), then redpress$(c_{it})<$redpress(c_{ju}).

(If the dissonance of c_{ju} is greater than the dissonance of c_{it}, then the attempt to reduce c_{ju} will be greater than the attempt to reduce c_{it})

According to Westermann, the **T**-theoretical terms of the Festinger theory are *pairdiss, paircons, diss,* and *redpress*, because these terms are determined by assuming the Festinger theory of cognitive dissonance itself. Therefore, its partial potential models are as follows:

DIII-3: y is a partial potential model of Festinger's theory of cognitive dissonance (y ∈ $\mathbf{M_{pp}}$(DissB)) iff there exists x such that:

(1) x = ⟨Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress⟩ ∈ $\mathbf{M_p}$(DissB)

(2) pairdiss, paircons, diss, redpress are **T**-theoretical.

(3) y = ⟨Time, Rawcog, Cognition, Disscog, Conscog, pairimp ⟩ ∈ $\mathbf{M_{pp}}$(DissB)

3.1.2.2. The Theory-Element of Forced Compliance Dissonance (DissF)

The potential models for the specialized theory element of forced compliance dissonance can be built by adding the following new components to the potential model of DissB (Westermann, 2000, pp. 203–204): Firstly, *Forcecom* is a subset of the cognition and pertain to the behaviors that are not in harmony with personal attitudes. Secondly, the function *attidiff* represents the difference between the real personal attitudes and the attitudes expressed in her behaviors. Thirdly, the function *imp* represents the subjective importance of cognition. Moreover, fourthly, the function *reward* shows the subjective magnitude of promised reward for the counter-attitudinal behavior or the magnitude of threatened punishment for refusing to do the counter-attitudinal behavior. This specialized theory element is called DissF, and its potential models can be defined as follows (Westermann, 2000, p. 203):

DIII-4: x is a potential model of the forced compliance dissonance (x ∈ $\mathbf{M_p}$(DissF)) iff there exist Time, Rawcog, Cognition, Disscog, Conscog,

pairdiss, paircons, pairimp, diss, redpress, Forcecom, attidiff, imp, reward such that:

(1) x = ⟨Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, attidiff, imp, reward⟩ ∈ $\mathbf{M_p}$(DissF)

(2) ⟨Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress⟩ ∈ $\mathbf{M_p}$(DissB)

(3) Forcecom ⊆ Cognition (cognitions on counterattitudinal behavior)

(4) attidiff: Forcecom → IR (attitudinal difference)

(5) imp: Cognition → IR_0^+ (importance of cognition)

(6) reward: Forcecom → IR_0^+ (magnitude of reward or punishment)

The actual models for DissF **M**(DissF) can be defined by assuming that the potential models $\mathbf{M_p}$(DissF) are held and by adding the following law or law-like statements: firstly, "the more important the opinions or the behavior involved, and the smaller the promised reward or threatened punishment, the greater is the magnitude of dissonance that is created" (Westermann, 2000, p. 204). Secondly, "pressure to reduce forced compliance dissonance may be manifested in a reduction of the importance or value of the behavior and opinion involved, an enhancement of the subjective magnitude of the promised reward or threatened punishment, and a change of private opinion in accordance with public behavior, i.e., in a smaller difference between real and expressed personal attitude" (Westermann, 2000, p. 204). The actual models of the specialization in the forced compliance dissonance are defined as follows (Westermann, 2000, p. 204):

DIII-5: x is an actual model of the forced compliance dissonance (x ∈ **M**(DissF)**)** iff there exist Time, Rawcog, Cognition, Disscog, Conscog,

pairdiss, paircons, pairimp, diss, redpress, Forcecom, attidiff, imp, reward such that:

(1) $\mathbf{x} = \langle$Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, attidiff, imp, reward$\rangle \in \mathbf{M_p}$(DissF)

(2) For all $c_{ju} \in$ Forcecom:
if_{cp} imp(c_{it}) < imp(c_{ju})
or reward(c_{it}) > reward(c_{ju})
then_{p} diss(c_{it}) < diss(c_{ju}).

(3) For all c_{it}, c_{ju}, c_{it+}, $c_{ju+} \in$ Forcecom with t<t+, u<u+:
if_{cp} 0 < redpress(c_{it}) < redpress(c_{ju})
then_{p} 0 > imp(c_{it+}) - imp(c_{it}) > imp(c_{ju+}) - imp(c_{ju})
or 0 < reward(c_{it+}) - reward(c_{it}) < reward(c_{ju+}) - reward(c_{ju})
or 0 > attidiff(c_{it+}) - attidiff(c_{it}) > attidiff(c_{ju+}) - attidiff(c_{ju}).

The additional terms are non-theoretical with respect to dissonance theory because their values can be determined by direct ratings, magnitude estimations, pair comparisons, or other standard scaling methods. Since DissF is a specialization of DissB, the **T**-theoretical terms of DissF are similar to those of DissB with respect to the theory-net according to DVIII-3 in BMS, p. 392. The partial potential models of DissF can be built by omitting its **T**-theoretical elements as follows:

DIII-6: y is a partial potential model of the forced compliance dissonance (y $\in \mathbf{M_{pp}}$(DissF)) iff there exists x such that:

(1) $\mathbf{x} = \langle$Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, attidiff, imp, reward$\rangle \in \mathbf{M_p}$(DissF)

(2) pairdiss, paircons, diss, redpress are **T**-theoretical.

(3) y = $\langle$Time, Rawcog, Cognition, Disscog, Conscog, pairimp, Forcecom, attidiff, imp, reward$\rangle \in \mathbf{M_{pp}}$(DissF)

3.2. Neuroscience

The second scientific field, whose theories will be modeled here, is neuroscience. In the recent development of cognitive science, there are many kinds of research dedicated to exploring the relationship between mind and body, especially the brain and neural systems. On the neurobiology level, mental processes depend on two primary processes, namely the information processed by the neural network and the chemical processes by the enzymes. The network process is how the neurons in their network process the data input to give appropriate responses. The chemical process is how enzymes or some chemicals play a role in making the network process more effective or less effective.

There are many types of research in both processes related to the theory of cognitive dissonance. Because of the time limitation, this dissertation will focus on the network process. Therefore, in this part of Chapter 3, two theory elements will be presented: the McCulloch-Pitts neuron and Hawkins-Kandel's computational neuroscientific theory (CNT). The author created the structuralist model of the McCulloch-Pitts neuron, whereas the structuralist model of Hawkins-Kandel's CNT was created by John Bickle – with some adaptation for this dissertation by the author.

3.2.1. Building a Structuralist Model of Hawkins-Kandel's Computational Neuroscientific Theory (CNT)

John Bickle builds a structuralist model of Hawkins-Kandel's Computational Neuroscientific Theory (CNT) based on their paper entitled *Is There a Cell-Biological Alphabet for Simple Forms of Learning?* (1984) In their paper, Hawkins and Kandel discuss the correlation between the phenomena of learning and the interneuron activities in the brain. Some progress has been made in identifying cellular mechanisms for habituation, sensitization, and conditioning in simple vertebrate systems and higher invertebrates such as *Aplysia*, *Drosophila*, *Hermissenda*, locust, and crayfish.

From these studies, Hawkins and Kandel (1984) summarize the general features of neural activities in the brain with respect to learning as follows: "(1) Elementary aspect of learning are not distributed in the brain but can be localized to the activity of specific nerve cells; (2) Learning produces alterations in the membrane properties and the synaptic connections of those cells; (3) The changes in synaptic connections so far encountered have not involved the formation of the new synaptic contacts. Rather, they are achieved by modulation the amount of chemical transmitter released by the presynaptic terminal of neurons; And (4) In several instances, the molecular mechanisms of learning involve intracellular second messengers and modulation of specific ion channels" (p. 375).

Bickle does not build the model of all these features, but only a part of them. He focuses on the connections among neurons in the learning process and broadens his approach to the general idea of mental representations. "[Mental] Representations are usually characterized in one of two ways: as patterns of activation values (values of non-negative real numbers representing the firing rates of some or all of the neurons in the network), or as patterns of synaptic weight values that regulate activation values in all but the input neurons of the network" (Bickle, 1998, p. 191). Bickle identifies the fundamental features of Hawkins and Kandel's neurobiological theories: "(1) a set of neurons, (2) a state of activation and an output function for each neurons, (3) a pattern of connectivity among the neurons (4) an intraneuronal activation rule for combining the inputs to a neuron with its current activation state to produce a new activation state, and (5) various intraneuronal processes for adjusting the synaptic strengths between a neuron and others receiving its output as part of their input" (Bickle, 1998, p. 191).

To build his structuralist model, Bickle uses the following notations: "*N* is a network (hence the well-ordering condition) of neurons (n). *Act* is a set of action commands (to the motor system). *T* is a set of time instances (t). *AV* is the activation-value relation, taking neurons at times into positive real

values. *O* is the neurons' output at times, determined by some mathematical function A on each neuron's activation value. *I* is [the] input at times coming into any neurons from outside the network (e.g., from the sensory periphery). *CW* is the connection weight relation at times, with the restriction that no unit is actively connected to itself. *Cause* is the causation relation from the activation values of a subset of neurons (seemingly without exception a proper subset, the output neurons) onto action commands. [Italic is used here to indicate the sets' names]" (Bickle, 1998, p. 192).

The potential models of the Computational Neuroscientific Theory ($\mathbf{M_p}$(CNT)) according to Bickle are defined as follows:

DIII-7: x is a potential model of the Computational Neuroscientific Theory ($x \in \mathbf{M_p}$(CNT)) iff there exist N, Act, T, AV, O, I, CW, Cause such that:

(1) $x = \langle N, Act, T, IN, IR, AV, O, I, CW, Cause \rangle \in \mathbf{M_p}$(CNT)

(2) N is a finite, non-empty, non-singleton, well-ordered set of neurons

(3) Act is a finite, possibly empty set of the action command

(4) T is a finite, non-empty set, non-singleton, well-ordered set of point of time

(5) $AV := N \times T \rightarrow IR^+$ (Activation of neurons)

(6) $O := N \times T \rightarrow IR$, and for all $n \in N$, $t \in T$, $O(n,t) = AV(n,t)$ (Neuron's Output)

(7) $I := N \times T \rightarrow IR$ (Neuron's Input)

(8) $CW := N \times N \times T \rightarrow IR$, & for all $n \in N$, $t \in T$, $CW(n,n,t) = \{x | x \in IR\}$ (Connection Weight at t)

(9) $Cause := AV^* \rightarrow Act$, where $AV^* \subseteq AV$ (Seemingly without exception, $AV^* \subset AV$)

The actual models contain the following laws or law-like statements as follows (Bickle, 1998, pp. 192–193): "(1) the activation value of each neuron at some time is the result of some arithmetical function *F* on

connection weights multiplied by the output of all neurons actively connected with the one in question, inputs to the neuron from outside the network [if any], and the neuron's activation value at the time instant just before. (2) If the right relation obtains between inputs to the network over time, then the connection weights times outputs of the presynaptic units differ by quantity *D* from those products at an earlier time. (3) if the same inputs to the network obtain over time, then the output of some units will be less at a later time compared to the earlier time." The set of the actual models of the Computational Neuroscientific Theory (**M**(CNT)) according to Bickle is defined as follows:

DIII-8: x is an actual model of the Computational Neuroscientific Theory (x $\in$**M**(CNT)) iff there exist N, Act, T, AV, O, I, CW, Cause such that:

(1) x = $\langle$ N, Act, T, IN, IR, AV, O, I, CW, Cause $\rangle \in$ $\mathbf{M_p}$(CNT)

(2) For all n_1, $n_2 \in$ N: AV(n_1, t) = CW(n_2, n_1) . O(n_2) I(n_1) AV(n_1, t-1).

(3) If Relation between I(N) at t & I(N) at t-1 &...& I(N) at t-n, then CW(N) at t times O(N) at t differs by D from CW(N) at t-n times O(N) at t-n).

(4) If I(N) at t= I(N) at t-1 = … = I(N) at t-n, then O(N) at t differs from O(N) at t-n in that for some n$\in$N, O(n,t) < O(n,t-n).

The partial potential models of computational neuroscientific theory can be defined by omitting the **T**-theoretical elements. Because all the elements can be observed empirically or be modeled by some other theory, the partial potential models of CNT are identical with its potential models. The partial potential models of the computational neuroscientific theory ($\mathbf{M_{pp}}$(CNT)) can be characterized as follows:

DIII-9: y is a partial potential model of the Computational Neuroscientific Theory (y $\in$ $\mathbf{M_{pp}}$(CNT)) iff there exists x such that:

(1) x = $\langle$ N, Act, T, IN, IR, AV, O, I, CW, Cause $\rangle \in$ $\mathbf{M_p}$(CNT)

(2) There is no **T**-theoretical element.

(3) $y = \langle N, Act, T, IN, IR, AV, O, I, CW, Cause \rangle \in \mathbf{M_{pp}}(CNT)$

3.2.2. Building a Structuralist Model of the McCulloch-Pitts Neuron

The McCulloch-Pitts neuron (1943) is an abstract and simplified model about the activity of a neuron in neural networks from neurophysiological data based on the following five assumptions: "1) The activity of the neuron is an "all-or-none" process. 2) A certain fixed number of synapses must be excited within the period of latent addition in order to excite a neuron at any time, and this number is independent of previous activity and position on the neuron. 3) The only significant delay within the nervous system is the synaptic delay. 4) The activity of any inhibitory synapse absolutely prevents excitation of the neuron at that time. 5) The structure of the net does not change with time" (McCulloch-Pitts, 1943, p. 118).

The McCulloch-Pitts model shows that a neuron has the characteristics of a digital automaton in its activities. A neuron consists of a soma and an axon. In a neural network, axons connect to other neurons at synapses. These synapses are a place where the information is transmitted from one neuron to another one. A neural network is an arrangement of a finite number of neurons, whereas every axon of a neuron is in connection with the soma of other neurons (or maybe its own) at the synapse. In a network these neurons have just two possible conditions – they fire (in excitation) or do not fire (in inhibition). These conditions can be achieved because each neuron has a threshold. If the amount of total input received by a neuron is higher than or the same as its threshold, the neuron is in an excitatory state; otherwise, the neuron is in an inhibitory state.

The threshold of neurons, according to the McCulloch-Pitts model, enables the neurons in a neural network to perform all logical operations. We can build various logical switches from such neurons by controlling the input

signal and the threshold: for a logical AND-switch, we choose the input signal 1/m and the threshold (m-1)/m, such that the neuron fires, if all inputs are active – m is here the number of input neurons. For a logical OR-switch, we choose the threshold 1/(m-1). Moreover, for the logical negation, we choose the threshold -1/2 and the synaptic weight -1, such that the threshold can be exceeded if the input signal is 0. A combination of these three operations can build the rest of the logical switches. Nowadays, this idea is developed and applied for various usages not only as an explanatory model in neuroscience but also in several other scientific fields, where the simulation of cognitive dissonance is just one among them.

The McCulloch-Pitts model and other models in artificial neural networks studies are typically presented as a directed graph. The vertices represent the neurons, whereas the edges represent the synaptic connections (axon and its synaptic connection). Every edge is labeled with a real number representing the strength of synaptic connections. A positive number indicates excitatory synapses, whereas the 0 (zero) or a negative number indicates the inhibitory synapses. For each neuron in the networks, there is a specific threshold value to fire. If the neuron's threshold value is surpassed at the time (t-1) by a single or several firing neurons connected to the soma of this neuron, the neuron will be in the excitatory condition and fires on time t.

In the McCulloch-Pitts model, the neurons (N) receive their input (Inp) from a number n of other neurons through a synaptic connection (C). The neurons that serve as input-giver are called "input neurons" (N_0). The synaptic connections between neurons have a synaptic weight (W), whereas every neuron has a threshold (θ). Suppose some input-neurons fire at time t_0. They give an input (Inp) for the neuron after them in a synaptic connection. As a response, a neuron will fire at time t_1, by giving its output ($Outp$). In the McCulloch-Pitts model, each neuron has only two conditions, namely inhibitory, represented by 0, and excitatory, represented by 1.

Three processes occur in the McCulloch-Pitts neuron. (1) The neuron unifies all inputs, that are received. This process forms the network-input for it. The process is represented here as a network-input function (*fnet*) as follows:

$$\text{fnet}\ (\text{Inp}, W, t_0) = \text{net}_n = \Sigma_{i=1}^{n}\, w_i.\text{inp}_i$$

(2) The second process is activation (*fact*). In this process, the network-input is compared with the neuron's threshold: if the network input is greater than or the same as the threshold, a neuron is in the excitatory condition; otherwise, a neuron is in the inhibitory condition.

$$\text{fact}\ (\text{net}_n, \theta) = \text{Act}_n = \begin{cases} 1, \text{ if } \text{net}_n \geq \theta \\ 0, \text{ otherwise} \end{cases}$$

(3) If the neuron is in the excitatory state, it will fire (output = 1) according to the output-function *fout*, otherwise, it will not fire (output = 0): *fout* (act_n) = output neuron at t_1.

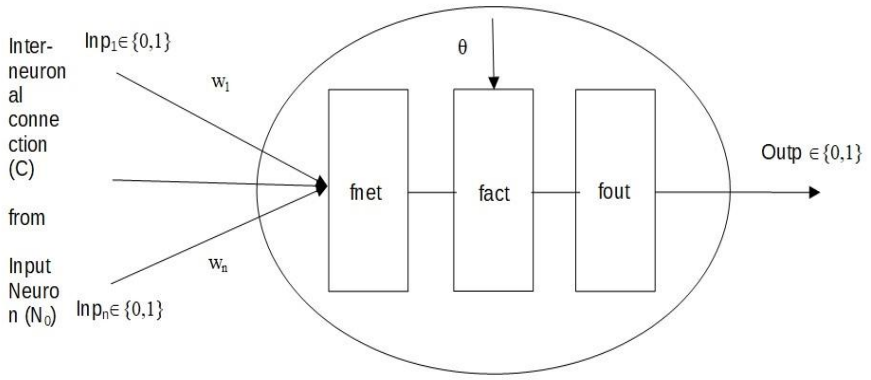


Figure 3.1. The McCulloch-Pitts model of a neuron (Adapted from Borgelt, C. et. al., 2003, p.33)

Based on the considerations discussed above, a theory-element of the McCulloch-Pitts neuron can be built as follows: Firstly, the potential models

of the McCulloch-Pitts model of a neuron ($\mathbf{M_p}$(MCP-N)) can be characterized as follows:

DIII-10: x is a potential model of McCulloch-Pitt model of neuron ($x \in \mathbf{M_p}$(MCP-N)) iff there exist N, N_0, T, IR, IN, C, W, θ, Inp, Outp, fnet, fact, fout, such that

(1) $x = \langle N, N_0, T, IR, IN, \theta, C, W, Inp, Outp, fnet, fact, fout \rangle \in \mathbf{M_p}$(MCP-N)

(2) N = a finite non-empty set of neurons

(3) N_0 = non-empty set of input neurons = $\{m \in N | (m, n) \in C\}$ (Input units)

(4) T = a discrete order of points of time t = 0, 1, 2, ... (Time)

(5) $\theta := N \rightarrow IR$

(Threshold – assigns to every neuron a real number as its threshold)

(6) $C \subseteq N \times N$

(a finite non-empty set of connection between neurons)

(7) $W := C \rightarrow IR$

(Synaptic Weight – assigns to each pair of neurons a real number as synaptic weight, where w(i,j) = w(j, i) and $w \in W$)

(8) $Inp := N/N_0 \times C \times T \rightarrow \{0,1\}$

(Input – assigns to each neuron, except input neurons, at the point of time *T* several real numbers as its input, that is sent by its input units (N_0) in the network; 0 = by an inhibitory input neuron and 1 = by an excitatory input neuron)

(9) $Outp := N/N_0 \times T \rightarrow \{0,1\}$

(Output – assigns to each neuron, except input neurons, at the point of time *T* a real number as its output, that is sent to the next neuron in the network; 0 = inhibitory and 1 = excitatory)

(10) $fnet := W \times Inp \rightarrow IR$

(Network Input function – assigns to each neuron (except input units) at every point of time t from *T* a real number as network input)

(11) fact:= fnet × θ → IR

(Activation function – compares the result from *fnet* with θ and assigns to the neuron a real number. The result is either 0 or 1)

(12) fout:= fact → Outp

(Output function – assigns every neuron a number 0 (inhibitory) or 1 (excitatory) as its output according to Outp.)

The actual models of McCulloch-Pitts model of neuron (**M**(MCP-N)) can be defined as follows:

DIII-11: x is an actual model of McCulloch-Pitts model of neuron ($x \in$ **M**(MCP-N)) iff there exist N, N_0, θ, C, W, T, Inp, Outp, fnet, fact, fout, IR, IN such that:

(1) $x = \langle N, N_0, T, IR, IN, \theta, C, W, Inp, Outp, fnet, fact, fout \rangle \in \mathbf{M_p}$(MCP-N)

(2) There is $n \in N/N_0$, $c_i \in C$ for $i \in IN$, $t_0, t_1 \in T$, and let net_n, act_n, out_n so that:

(2.1) $net_n = fnet\ (Inp, W, t_0) = \sum_{i=1}^{n} Inp_i\ (n, c_i, t_0).W(c_i)$

(2.2) $act_n = fact(\ net_n, \theta)$:

(i) $act_n = 1$, if $net_n \geq \theta$,

(ii) $act_n = 0$, otherwise.

(2.3) $out_n = fout\ (act_n) = Outp\ (n, t_1)$.

Now we define the partial potential models of the McCulloch-Pitts neuron by omitting the **T**-theoretical concepts. In the McCulloch-Pitts model, only the three function terms – *fnet, fact,* and *fout* – are **T**-theoretical because the concepts of the neuron, connections, synaptic weight, time, and threshold are empirical. These three terms are **T**-theoretical because these terms presuppose this theory of neuron itself. According to this theory a neuron must have the following three characteristics: First, receiving input-signals

from the sending neurons, called the input neurons, as represented by *fnet*. The second characteristic is comparing the input signal with a specific weight called 'threshold.' The neuron will fire if the input signal is greater than its threshold (excitatory state); the neuron will not fire if the input signal is smaller than its threshold (inhibitory state). *fact* represents this characteristic. Moreover, *fout* represents the third characteristic, namely that of sending the result to other receiving neurons. Therefore, the partial potential models of McCulloch-Pitts neuron are characterized as follows:

DIII-12: y is a partial potential model of McCulloch-Pitts neuron ($\mathbf{M_{pp}}$(MCP-N)) iff there exists x such that:

(1) $x = \langle N, N_0, T, IR, IN, \theta, C, W, Inp, Outp, fnet, fact, fout \rangle \in \mathbf{M_p}$(MCP-N)

(2) fnet, fact, fout are **T**-theoretical.

(3) $y = \langle N, N_0, T, IR, IN, \theta, C, W, Inp, Outp \rangle \in \mathbf{M_{pp}}$(MCP-N)

3.3. The Artificial Neural Network

Today it is still impossible to know exactly how the network of neurons in the brain works produce the phenomena of cognition because of its complexity. In neuroscience, we can only learn about the parts of the brain on various levels of explanation that play an essential role in cognitive processes, such as neurons and their network, regions in the cerebrum, the interconnection between parts of the brain and so on. Our brain has 10^{11} neurons, and therefore, there are $n.2^n$ possibilities of connections among neurons, where $n = 10^{11}$; we have no chance to fully understand the cognitive process in the brain with our current scientific development.

Many neuroscientists and psychologists seek another way out to understand the brain's information processing from a branch of computer science, called Artificial Intelligence (A.I.). In the history of artificial intelligence, there are two approaches developed to modeling cognition

namely, by symbol manipulation and by the artificial neural network (also known as connectionism) (Bechtel and Abrahamsen, 2002, p. 2): "Both connectionist and symbolic systems can be viewed as computational systems. However, they advance quite different conceptions of what computation involves. In the symbol approach, computation involves the transformation of symbols according to [logical] rules. … We treat a traditional computer as a symbolic device, and we view it as performing symbolic manipulations specified by rules which typically are written in a special data structure called the *program*. [Italics in the original] The connectionist view of computation is quite different. It focuses on causal processes by which units excite and inhibit each other and does not provide either for stored symbols or rules that govern their manipulation." The symbolic manipulation approach, whose proponents are Dennett, Fodor, Pylyshyn, and others, has its roots more in logic and linguistics, whereas the connectionist approach is inspired by a model of a neuron from neuroscience and statistics or probability theory. Its proponents are Frank Rosenblatt, John Hopfield, Geoffrey Hinton, David Rummelhart, Paul Smolensky, David McClelland, among others. Because of this close relation between neuroscience and the artificial neural network, I have chosen to use the artificial neural network as a model, or respectively a simulation of cognitive dissonance reduction.

Giving an explanation via a simulation of the brain's computational process is just one of the goals of the artificial neural network research. The other goal is to solve some technical issues, such as face recognition, controlling, voice recognition, and so on. This dissertation will limit itself to the first goal, i.e., to analyze the intertheoretical connections between Festinger's theory of cognitive dissonance and the artificial neural network by the simulations of cognitive dissonance according to the following models:

(1) Thomas R. Shultz and Mark R. Lepper: the consonance model.

(2) Frank van Overwalle and Karen Jordens: the adaptive connectionist model.

The abbreviation ANN stands for the artificial neural network.

In the simulation of psychological phenomena by using ANN, they built a specific network of several artificial models of the neuron – normally the Rosenblatt perceptron. These artificial neurons are placed in a certain network pattern, such as a feed-forward network or a recurrent network. Then specific learning algorithms (or learning rules), such as the delta rule (also called Widrow-Hoff rule), will be executed. Whereas the consonance model uses the Hopfield network – a kind of recurrent network, the adaptive connectionist model – to simplify it will be called the connectionist model – uses the two layers feed-forward neural network and the delta rule as its learning algorithm. In this third section of Chapter 3, we will build several structuralist models for the Rosenblatt perceptron, the Hopfield network, the two-layers feed-forward neural network, and the delta rule.

3.3.1. Building a Structuralist Model of the Rosenblatt Perceptron

The McCulloch-Pitts model has several limitations. Firstly, the input from input neurons is only either 1 or 0. Secondly, the connection weights and the threshold are initially set from the beginning to perform certain logical functions. Therefore, the McCulloch-Pitts neuron cannot learn. In history, these limitations were removed by the theory of the perceptron due to Arthur Rosenblatt in 1958.

The Rosenblatt perceptron is the second generation of the model of a neuron. In the perceptron, the input(s) are not only 0 and 1, as in the McCulloch-Pitt neuron. The input can be various numbers depending on the use of the network. Therefore, instead of just having the characteristics of a digital automaton, the perceptron has a statistical character. It can now analyze a given set of data and build an approximative model for those data by its activation function (and learning rule).

Perceptrons (or neurons), which are not input-perceptrons, (N/N_0) receive input-value in real numbers from input-perceptrons (or also input

neurons) (N_0) through a (synaptic) connection (C). The synaptic connections between neurons have a synaptic weight (W), whereas every neuron, which is not input neuron, has a bias (B) that is normally set as 1. The notion of bias here is like the notion of threshold in the McCulloch-Pitts neuron. By setting bias =1, the bias gives the perceptron a trainable constant value, because the connection weight of bias can be adjusted according to the learning rule. By implementing a learning rule, such as the delta rule, the bias helps to adjust the function to approximate the data. The synaptic weight of the connection between the bias and the neuron also has a weight W_0. Because of bias and its connection, the perceptron has the characteristics of the statistical function in its network-input function.

Like the McCulloch-Pitts neuron, the perceptron also has three processes. (1) The neuron unifies all inputs that are received. It forms the network-input for it. This process is represented here as a network function (*fnet*) as follows:

$$\text{fnet (Inp, W)} = \text{net}_n = \sum_{i=1}^{n} \text{Inp}_i\,(n, c_i,).W(c_i).$$

(2) The second process is the activation function (*fact*). In this process the network-input is processed according to a certain activation function.[1] The activation function that will be used in the connectionist model in Chapter 7 is the linear regression:

$$\text{fact}(\text{net}_n, b, w_0) = \text{act}_n = \text{net}_n + b.w_0$$

[1] For neurons in the hidden layer one of the following activation functions is normally applied:

a. The sigmoid activation function:

$$\text{fact}(\text{net}_n, b, w_0) = \text{act}_n = 1/ (1 + e^{\wedge}(\text{net}_n + b.w_0))$$

b. The Tanh activation function:

$$\text{fact}(\text{net}_n, b, w_0) = \text{act}_n = (e^{\wedge}(\text{net}_n + b.w_0) - e^{\wedge}\text{-}(\text{net}_n + b.w_0))/ (e^{\wedge}(\text{net}_n + b.w_0) + e^{\wedge}\text{-}(\text{net}_n + b.w_0))$$

c. The Rectified linear (ReLU) activation function:

$$\text{fact}(\text{net}_n, b, w_0) = \text{act}_n = \begin{cases} 0 & \text{for } (\text{net}_n + b.w_0) < 0 \\ \text{or} & \\ (\text{net}_n + b.w_0) & \text{for } (\text{net}_n + b.w_0) >= 0 \end{cases}$$

And the neurons in the output layer normally use linear regression:

$$\text{fact}(\text{net}_n, b, w_0) = \text{act}_n = \text{net}_n + b.w_0$$

(3) The neuron gives a real number as its output according to:

$$fout(act_n) = out_n.$$

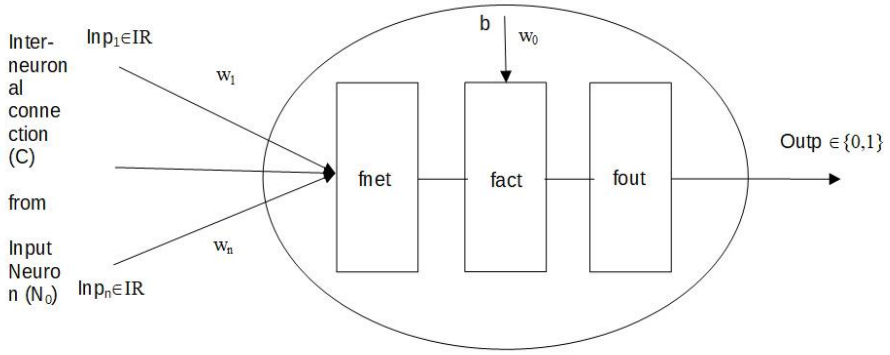


Figure 3.2. The Rosenblatt model of a perceptron (Adapted from Borgelt, C. et. al., 2003, p.33)

From the description above, we can build a structuralist model of the Rosenblatt perceptron according to the following steps: First, we define the potential models of the Rosenblatt perceptron ($\mathbf{M_p}$(RP)) as follows:

DIII-13: x is a potential model of the Rosenblatt perceptron ($x \in \mathbf{M_p}$(RP)) iff there exist N, N_0, IR, IN, B, C, W_0, W, Inp, Outp, fnet, fact, fout, such that:

(1) $x=\langle N, N_0, IR, IN, B, C, W_0, W, Inp, Outp, fnet, fact, fout\rangle \in \mathbf{M_p}(RP)$

(2) N = a finite non-empty set of neurons

(3) N_0 = a non-empty set of input-neurons = $\{m \in N | (m,n) \in C\}$
(Input units)

(4) $B := N/N_0 \rightarrow IR$
(Bias – assigns to every neuron besides the input neurons a real number as its bias. Bias is normally set = 1)

(5) $C \subseteq N \times N$
(a finite non-empty set of connections between neurons)

(6) $W_0 := B \times N/N_0 \rightarrow IR$ (Synaptic Weight from Bias)

(7) $W := C \rightarrow IR$

(Synaptic Weight – *W* assigns to each pair of neurons a real number as the synaptic weight.)

(8) Inp:= $N/N_0 \times C \rightarrow IR$

(Input – assigns to each neuron several real numbers as its input, sent by its input-units in the network)

(9) Outp:= $N/N_0 \rightarrow IR$

(Output – assigns to each neuron a real number as its output, that is sent to the next neuron in the network)

(10) fnet: $W \times Inp \rightarrow IR$

(Network Input function – assigns to neurons (except input units) a real number as the network input)

(11) fact: $fnet \times \beta \times W_0 \rightarrow IR$

(Activation function – there are various activation functions)

(12) fout: $fact \rightarrow Outp$

(Output function – assigns every neuron a real number as its output according to Outp)

Second, the actual models of the Rosenblatt perceptron (**M**(RP)) can be defined as follows:

DIII-14: x is an actual model of the Rosenblatt perceptron ($x \in$ **M**(RP)) iff there exist N, N_0, IR, IN, B, C, W_0, W, Inp, Outp, fnet, fact, fout such that.

(1) $x=\langle N, N_0, IR, IN, B, C, W_0, W, Inp, Outp, fnet, fact, fout\rangle \in \mathbf{M_p}(RP)$

(2) There is $n \in N/N_0$, $c_i \in C$ for $i \in IN$, $b \in B$ and let net_n, act_n, out_n so that:

(2.1) $net_n = fnet\ (Inp, W) = \sum_{i=1}^{n} Inp_i\ (n, c_i,).W(c_i)$,

(2.2) $act_n = fact(net_n, b, w_0) = net_n + b.w_0$.

(2.3) $out_n = fout\ (act_n) = Outp$.

Finally, the partial potential models of the Rosenblatt perceptron ($\mathbf{M_{pp}}$(RP)) can be defined by omitting the **T**-theoretical elements. Like in the

McCulloch-Pitts model, only *fnet, fact,* and *fout* of the Rosenblatt perceptron are **T**-theoretical because the other concepts, such as the neuron, bias, connections, bias's connection weight, and synaptic weight are 'empirical' – determined by inputs containing data from an empirical observation – and based on other theories. These three terms are **T**-theoretical because these terms presuppose this theory of a perceptron itself. Later in Chapter 6 we can see that these three are theoretical terms related to the net with McCulloch-Pitts neuron. The partial potential model of the Rosenblatt perceptron can be defined as follows:

DIII-15: y is a partial potential model of the Rosenblatt perceptron ($\mathbf{M_{pp}}$(RP)) iff there exists x such that:

(1) $x = \langle N, N_0, IR, IN, B, C, W_0, W, Inp, Outp, fnet, fact, fout \rangle \in \mathbf{M_p}(RP)$.

(2) fnet, fact, fout are **T**-theoretic.

(3) $y = \langle N, N_0, IR, IN, W, B, W_0, Inp, Outp \rangle \in \mathbf{M_{pp}}(RP)$.

Before discussing the architecture of the network, the following items are worthy of consideration: (1) There are two main streams of the artificial neural network regarding the goal and application of its development. The first develops the artificial neural network to build the simulation of the brain's functionality, and the second develops the artificial neural network to solve a specific problem or support some technology as a kind of artificial intelligence or machine learning. The first one tries to mimic how the brain works as precisely as possible. Therefore, it usually uses the perceptron with the input and output like the McCulloch-Pitts model – 0 for the inhibitory condition and 1 for the excitatory condition. As for the second idea, the perceptron model is generally used with various inputs and outputs in the real numbers. Because of its purposes, this dissertation will follow the first idea. (2) For a multi-layers feed-forward neural network with hidden layers, the model of the perceptron is usually used. For the simplest version with only

two layers – input-layer and output-layer – both the McCulloch-Pitts model and the Rosenblatt perceptron can be used.

3.3.2. Building a Structuralist Model of the Network Architecture

The perceptrons are placed in a specific architecture of networks in order to work according to our purpose. There are various network-architectures in the study of artificial neural networks. However, we can categorize them into two main categories: feed-forward neural networks and recurrent neural networks. (People may also build a mixture of them).

A network-architecture of neural networks is normally described in the terms of directed graph-theory. As Borgelt et.al., 2003, pp. 29–30 say: "*Ein (künstliches) neuronales Netz ist ein (gerichteter) Graph G = (U,C), dessen Knoten u ∈ U Neuronen (neurons, units) und dessen Kanten c ∈ C Verbindungen (connections) heißen. Die Menge U der Knoten ist unterteilt in die Menge U_{in} der Eingabeneuronen (input neurons), U_{out} der Ausgabeneuronen (output neurons) und die Menge Uhidden der versteckten Neuronen (hidden neurons). Es gilt* [In English: An (artificial) neural network is a (directed) graph G = (U, C) whose nodes are called u ∈ U neurons (neurons, units) and whose edges are c ∈ C connections. The set U of nodes is subdivided into the set U_{in} of the input neurons, the set U_{out} of the output neurons, and the set U_{hidden} of the hidden neurons. It applies]"*:*

$$U = U_{in} \cup U_{out} \cup U_{hidden},$$

$$U_{in} \neq \emptyset \qquad U_{out} \neq \emptyset \qquad U_{hidden} \cap (U_{in} \cup U_{out}) = \emptyset.$$"

In this work, I will use notation *N* for Neuron, instead of *U*.

In an artificial neural network-architecture, these three kinds of neurons-layers have each of their roles as follows: (1) Input neurons (N_{in}) are the neurons in the input layer. They receive the input values for the neural network and convey them to the neurons in the next layers, either the output layer (in two-layer neural networks) or the first hidden layer (in multi-layer

neural networks). (2) Output neurons (N_{out}) are the neurons in the output layer. They receive the values, that were processed and transferred by the other layers, either the last hidden layer or the input layer, and give them back as the output of the networks after processing them. (3) Hidden neurons (N_{hidden}) are the neurons in the hidden layer. Hidden layers are the layers of neurons that are neither the input layer nor the output layer. Hidden layers lie between the input layer and the output layer – they are called "hidden layer" because of their place. They received the output of the neuron in the previous layer, process it, and convey the result as an input to the next layer.

Between two neurons, which are connected by a directed connection *C*, we can define the neuron predecessors (*pred*) and the neuron successors (*succ*) as follows (Borgelt et al., 2003, p. 29):

$$pred = \{n_1 \in N | (n_1, n_2) \in C\}$$

$$succ = \{n_2 \in N | (n_1, n_2) \in C\}$$

In the neural network "*Jeder Verbindung* $(v,u) \in C$ *ist ein Gewicht* w_{uv} *zugeordnet und jedem Neuron* $u \in U$ *drei (reellwertige) Zuständsgrößen: die Netzeingabe* net_u *(network input), die Aktivierung* act_u *(activation) und die Ausgabe* out_u *(output). Jedes Eingabeneuron* $u \in U$ *in besitzt außerdem eine vierte (reellwertige) Zustandsgröße, die externe Eingabe* ext_u *(external input*) [in English: Each connection $(v, u) \in C$ is assigned a weight w_{uv} and each neuron $u \in U$ three (real-valued) states of operation: the network input net_u, the activation act_u (activation) and the output out_u (output). Each input neuron $u \in U$ in also has a fourth (real-valued) state variable, the external input ext_u (external input).]" (Borgelt, et.al., 2003, p.30). In this work, the notations n_1, n_2 is used here instead of u, v.

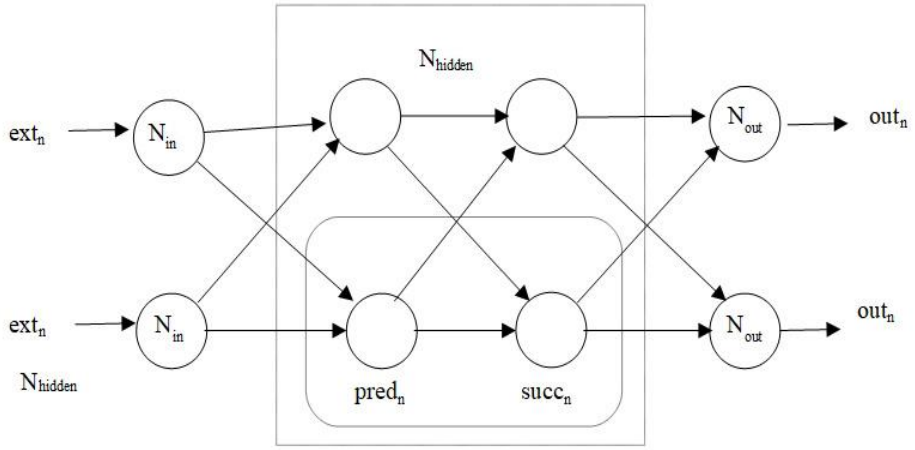


Figure 3.3. The architecture of an artificial neural network

To build a structuralist model for a theory element of the architecture of the artificial neural network, we must define the potential model, the actual model, and the partial potential model. The potential models of the architecture of the artificial neural network ($\mathbf{M_p}$(archNN)) can be characterized as follows:

DIII-16: x is a potential model of the architecture of the artificial neural network ($x \in \mathbf{M_p}$(archNN)) iff there exist N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n, net_n, act_n, out_n such that:

(1) $x = \langle N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n, net_n, act_n, out_n \rangle \in \mathbf{M_p}$(archNN)

(2) N is a finite non-empty set of neurons.

(3) N_{in} is a finite non-empty set of input-neurons.

(4) N_{out} is a finite non-empty set of output-neurons.

(5) N_{hidden} is a finite set of hidden neurons.

(6) $C \subseteq N \times N$ (a finite non-empty set of directed connections between neurons)

(7) $pred = \{n_1 \in N | (n_1, n_2) \in C\}$ (presynaptic neurons)

(8) $succ = \{n_2 \in N | (n_1, n_2) \in C\}$ (postsynaptic neurons)

(9) $W := C \rightarrow IR$ (synaptic weight)

(10) $ext_n := N_{in} \rightarrow IR$ (external input)

(11) $net_n := N \rightarrow IR$ (network-input)

(12) $act_n := N \rightarrow IR$ (activation)

(13) $out_n := N \rightarrow IR$ (output)

The following law or law-like statement determines the set of actual models of the architecture of artificial neural network (**M**(archNN)): (1) A network lets no single neuron excluded from the rest. All neurons are connected and play a role as input neurons, hidden neurons, or output neurons. (2) Input neurons play the role of predecessors in the network, whereas the output neurons play the role of successors, and the hidden neurons play both roles as successors and as predecessors. (3) Neurons in the hidden layer(s) are neither the input neurons nor the output neurons. The actual models for the architecture of artificial neural network (**M**(archNN)) can be formally characterized as follows:

DIII-17: x is an actual model of the architecture of the artificial neural network ($x \in$ **M**(archNN)) iff there exist N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n, net_n, act_n, out_n such that:

(1) $x = \langle N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n, net_n, act_n, out_n \rangle \in$ $\mathbf{M_p}$(archNN)

(2) $N = N_{in} \cup N_{out} \cup N_{hidden}$,

(3) for all $n \in N$ it holds:

(3.1) $N_{in} = pred$

(3.2) $N_{out} = succ$

(3.3) $N_{hidden} = succ \cap pred$

(4) $N_{hidden} \cap (N_{in} \cup N_{out}) = \varnothing$

(5) $\forall n \in N_{hidden}, \forall n \in N_{out}$:

(5.1) $net_n = f_{net}^{(n)} (\vec{w}_n, \vec{in}_n) = \vec{w}_n \vec{in}_n = \Sigma_{n0 \in pred(n)} w_{n0,n} out_{n0}$.

(5.2) For the Rosenblatt perceptron with the linear activation function:

$act_n = fact(net_n, b, w_0) = net_n + b.w_0$

(5.3) $out_n = fout\ (act_n)$

To characterize the partial potential models of the architecture of an artificial neural network (**M**$_{pp}$(archNN)), we omit the **T**-theoretical concepts from the potential models (**M**$_{p}$(archNN)). In the architecture of a neural network, the three terms – net_n, act_n, out_n – are **T**-theoretical, because these terms presuppose the concept of a network of neurons. They are the results of the three functions of neurons in the network – either according to the McCulloch-Pitts model or the Rosenblatt perceptron. The partial potential models of the architecture of a neural network (**M**$_{pp}$(archNN)) are characterized as follows:

DIII-18: y is a partial potential model of the architecture of neural network (y $\in$ **M**$_{pp}$(archNN)) iff there exist x such that:

(1) $x = \langle N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n, net_n, act_n, out_n \rangle \in$ **M**$_{p}$(archNN)

(2) net_n, act_n, out_n are **T**-theoretic.

(3) $y = \langle N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n \rangle \in$ **M**$_{pp}$(archNN)

3.3.3. Building a Structuralist Model for the Two-Layers Feed-Forward Neural Network and the Hopfield Network

The two network architectures that we will use, i.e., the two layers feed-forward neural network and the Hopfield network, are understood as two among many specializations of the general model of a network architecture according to the concept of theory-net because they can be derived by adding several additional requirements.

3.3.3.1. The Theory-Element of the Two Layers Feed-forward Neural Network

In feed-forward neural networks, the connections between neurons do not form a cycle. Through these connections, all neurons in a specific layer always send their outputs to neurons in the next layer in the direction from input to output. The neurons in the input layer send their output to the neurons in the output layer or the first layer of the hidden layer. The neurons in the hidden layer send their output to the neurons in the output layer or to the neurons in the next hidden layer. The neurons in the output layer receive the outputs of the neurons in the input layer or in the last hidden layer as their inputs.

From this scenario, we can see that there are two kinds of feed-forward neural networks. The simplest kind is called "two-layers feed-forward neural network," consisting of only two layers of neurons, namely a layer of input neurons (input layer) and a layer of output neurons (output layer). The input is fed directly by the input neuron to the neurons in the output layer because this kind of neural network has no hidden layer.

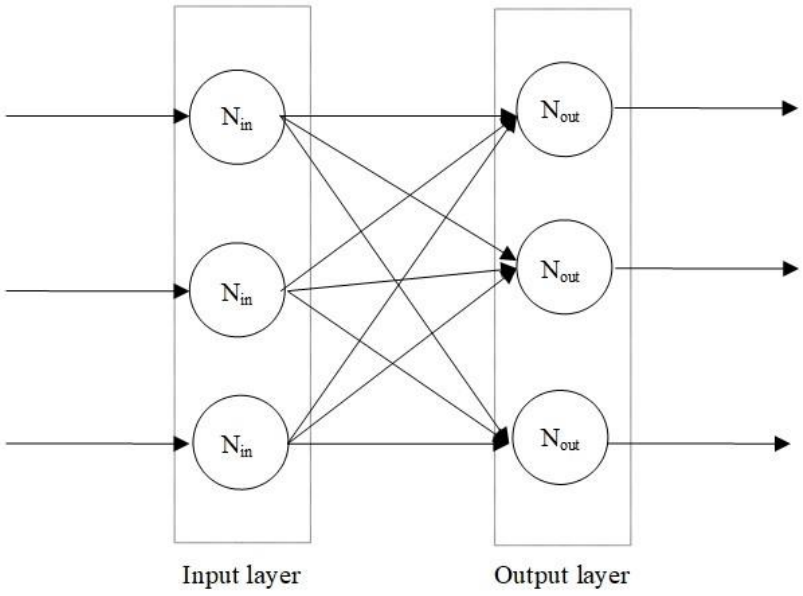


Figure 3.4. A two-layers feed-forward neural network

For this specialization, we add the following new laws statements (or law-like statements) in the actual models of two-layers feed-forward neural networks (**M**(2L-FFNN)), whereas the potential models of two-layers feed-forward neural networks (**M**$_{\mathbf{p}}$(2L-FFNN)) are identical with **M**$_{\mathbf{p}}$(archNN): (1) There is no hidden layer of neurons; therefore each neuron is either an input neuron or output neuron. (2) The input neurons and the output neurons are not identical. (3) All connections in this architecture are connections between input neurons and output neurons. (4) For each output neuron, its network input (net_n) is the result of the network input function of the neuron. Whereas its activation (act_n) is the result of the activation function of the neuron, and its network output is the result of the output function of the neurons. The actual models of two-layers feed-forward neural network-architecture (**M**(2L-FFNN)) can be characterized as follows:

DIII-19: x is an actual model of the two-layer feed-forward neural network (x $\in$ **M**(2L-FFNN)) iff there exist N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n, net_n, act_n, out_n such that:

(1) $x = \langle N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n, net_n, act_n, out_n \rangle \in$ **M**$_{\mathbf{p}}$(2L-FFNN)

(2) $N_{hidden} = \varnothing$

(3) $N = N_{in} \cup N_{out} \cup N_{hidden}$,

(4) for all $n \in N$ it holds:

(4.1) N_{in} = pred

(4.2) N_{out} = succ

(5) $N_{in} \cap N_{out} = \varnothing$

(6) $C \subseteq N_{in} \times N_{out}$

(7) $\forall n \in N_{out}$:

(7.1) $net_n = f_{net}^{(n)}(\vec{w}_n, \vec{in}_n) = \vec{w}_n \vec{in}_n = \Sigma_{n0 \in pred(n)} w_{n0,n} out_{n0}$.

(7.2) For perceptron with the linear activation function:

$act_n = fact(net_n, b, w0) = net_n + b.w0$

(7.3) $out_n = fout\ (act_n)$

The second type of feed-forward neural network is the multi-layers one. This kind consists not only of an input layer and an output layer but also some layers of hidden neurons (hidden layer) – at least one hidden layer. We do not discuss this type because it will not be used for the simulations discussed later.

3.3.3.2. A Theory-Element of the Hopfield Network

In the recurrent neural networks, the connections between neurons can form a loop. There are several architectures of the recurrent neural network, such as the Hopfield network, Boltzmann machine, etc. For the goal of this dissertation, only a structuralist model of the Hopfield network will be built.

The Hopfield network is a form of a fully recurrent neural network popularized by John Hopfield in 1982 through his paper *Neural Networks and Physical Systems with Emergent Collective Computational Abilities*. The neurons in the Hopfield network are binary threshold units in strong backward coupling.[2] They take only two different values for their states, 0 (inhibitory)

---

[2] "The processing devices will be called neurons. Each neuron i has two states like those of McCulloch and Pitts: $V_i = 0$ ("not firing") and $V_i = 1$ ("firing at maximum rate"). When neuron i has a connection made to it from neuron j, the strength of connection is defined as T_{ij}. (Nonconnected neurons have $T_{ij} = 0$.) The instantaneous state of the system is specified by listing the N values of Vi, so it is represented by a binary word of N bits.

The state changes in time according to the following algorithm. For each neuron i there is a fixed threshold U_i. Each neuron i readjusts its state randomly in time but with a mean attempt rate W, setting

$$V_i \to 0 \quad \text{If } \Sigma_{j \neq i} T_{ij} V_i > U_i$$

$$V_i \to 1 \quad \text{If } \Sigma_{j \neq i} T_{ij} V_i < U_i$$

Thus, each neuron randomly and asynchronously evaluates whether it is above or below threshold and readjusts accordingly. (Unless otherwise stated, we choose $U_i = 0$.)

Although this model has superficial similarities to the perceptron, the essential differences are responsible for the new results. First, perceptrons were modeled chiefly with neural connections in a "forward" direction A→B→C→D. The analysis of networks with strong backward coupling  proved intractable. All our interesting results arise as

or 1 (excitatory), so the Hopfield network uses a kind of McCulloch-Pitt neuron. To build a structuralist model of the Hopfield network, we can use *fnet*, *fact*, and *fout* of the McCulloch-Pitts neuron for this network.

The network input function of each neuron n is the sum of all outputs of other neurons [plus the actual input] times the connection's weight. In matrix form, it can be described as follows:

$$\forall n \in N: fnet^{(n)} (\vec{w}_n, \vec{in}_n) = \vec{w}_n \vec{in}_n = \Sigma_{m \in N-\{n\}} w_{mn} out_m.$$

The activation function of each neuron n is a threshold-function as follows:

$$\forall n \in N: fact^{(n)} (net_n, n) = \begin{cases} 1, \text{ in case } net_n \geq \theta_n, \\ -1, \text{ otherwise.} \end{cases}$$

Or sometimes the activation function of the neurons of a Hopfield network is defined by using the old activation act_n (Borgelt, C. et. al., 2003, p. 112):

$$\forall n \in N: fact^{(n)} (net_n, n, act_n) = \begin{cases} 1, & \text{in case } net_n > \theta_n, \\ -1, & \text{in case } net_n < \theta_n, \\ act_n, & \text{in case } net_n = \theta_n. \end{cases}$$

The output function of each neuron is the following.

$$\forall n \in N: fout^{(n)} (act_n) = act_n.$$

As a fully recurrent neural network, each neuron in the Hopfield network is connected to other neurons, except with itself. In the Hopfield network, every

consequences of the strong back-coupling. Second, Perceptron studies usually made a random net of neurons deal directly with areal physical world and did not ask the questions essential to finding the more abstract emergent computational properties. Finally, Perceptron modeling required synchronous neurons like a conventional digital computer. There is no evidence for such global synchrony and, given the delays of nerve signal propagation, there would be no way to use global synchrony effectively. Chiefly computational properties which can exist in spite of asynchrony have interesting implications in biology" (Hopfield, 1982, p. 2554).

neuron serves both as an input neuron and as an output neuron. Because all neurons in the Hopfield networks serve as input and output, it has no hidden neurons. We need to add the following statements in our general model of a neural network above:

(i) $N_{hidden} = \emptyset$, $N_{in}=N_{out}=N$,

(ii) $C = N \times N - \{(n, n) | n \in N\}$.

The weight of each connection is symmetrical.

$n_0, n_1 \in N, n_0 \neq n_1: w_{n0n1} = w_{n1n0}$.

This network has its convergence statement. Therefore, the Hopfield network does not need any learning rule, unlike the feed-forward neural network. A convergence statement is a law or law-like statement that is connected to the learning process of the neural network. In the learning process, the activation of neurons in the Hopfield networks are newly and asynchronously calculated. After many finite steps (max. $n.2^n$ steps of a single realization, with n = number of neurons), it will reach a stable condition, when the Hopfield network reaches one of the 'lowest' cost (or to converge to a local minimum). The stable condition is called "convergence."

As a function to reach a stable condition, the Hopfield network uses a so-called energy function. This energy function assigns to every state of the Hopfield network a real number as the energy of state. This function must become smaller or stay the same in every state-transition. The lowest energy defines the stable condition/state. The energy function of the Hopfield network is

$$E = - 1/2 \, act^{\rightarrow T} W \, act^{\rightarrow} + \theta^{\rightarrow T} act^{\rightarrow}$$

with $act^{\rightarrow} = (act_{u1}, \ldots, act_{un})^T$ = the activation states of nets.

W = the weight-matrix of nets

$\theta^{\rightarrow} = (\theta_{u1}, .., \theta_{un})^T$ is the vector of the threshold of neurons.

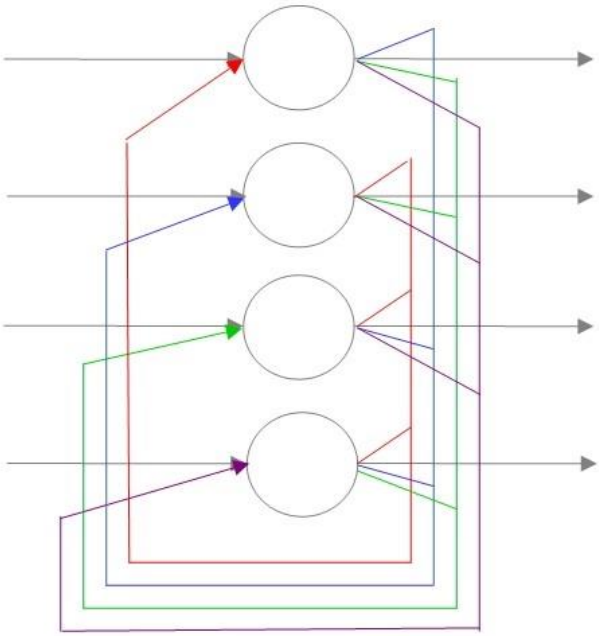


Figure 3.5. The Hopfield network with four neurons (Adapted from: Wikipedia)

To build a structuralist model of the Hopfield network, we must first modify the potential model by adding two extra concepts, namely state (*State*) and energy (*E*). The concept of state (*State*) is a three tuples relation between activation (act_n) and connection weights (*W*) and activation (act_n). Moreover, the energy (*E*) is a function mapping each state to a specific rational number. We can characterize the potential model of the Hopfield network ($\mathbf{M_p}$(HN))as follows:

DIII-20: x is a potential model of the architecture of the Hopfield network (x $\in$ $\mathbf{M_p}$(HN)) iff there exist N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n, net_n, act_n, out_n, State, E so that:

(1) x=$\langle$ N, N_{in}, N_{out}, N_{hidden}, IR, θ, C, pred, succ, W, ext_n, net_n, act_n, out_n, State, E $\rangle$ $\in$ $\mathbf{M_p}$(HN)

(2) N is a finite non-empty set of neurons.

(3) C $\subseteq$ N×N is a finite non-empty set of directed connections between neurons.

(4) N_{in} is a finite non-empty set of input neurons.

(5) N_{out} is a finite non-empty set of output neurons.

(6) N_{hidden} is a finite set of hidden neurons.

(7)	$\theta := N \rightarrow IR$	(Threshold/bias)
(8)	$pred(n) = \{n_1 \in N \mid (n_1, n_2) \in C\}$	(presynaptic neurons)
(9)	$succ(n) = \{n_2 \in N \mid (n_1, n_2) \in C\}$	(postsynaptic neurons)
(10)	$W := C \rightarrow IR$	(synaptic weight)
(11)	$ext_n := N_{in} \rightarrow IR$	(external input)
(12)	$net_n := Inp \times W \rightarrow IR$	(network input)
(13)	$act_n := N \rightarrow IR$	(activation)
(14)	$out_n := N \rightarrow IR$	(output)
(15)	$State \subseteq act_n \times W \times act_n$	(a finite non-empty set of states of Hopfield's net)
(16)	$E := State \rightarrow IR$	(Energy function)

The second step is the modification of the actual models of the neural network architecture. The Hopfield network can be seen as a specialization of the standard model by adding the following laws or law-like statement in the actual model:

(1) In the Hopfield network, there is no hidden layer.

(2) Input neurons and output neurons are identical. All neurons in a Hopfield's network serve both as input neurons and as output neurons.

(3) Every neuron is connected to all other neurons, except with itself.

(4) The connection weight between n_1 and n_0 is identical to the connection weight between n_0 and n_1.

(5) Each neuron network input (net_n) is the result of each neuron's input function (see above).

(6) For each neuron, activation (act_n) is the result of each neuron's activation function (see above).

(7) Each neuron network output (out_n) is the result of each neuron's output function (see above).

(8) Energy function: $E = - 1/2\ act^{\rightarrow T}\ W\ act^{\rightarrow} + \theta^{\rightarrow T}\ act^{\rightarrow}$ (for the details see above).

The actual models for the Hopfield network (**M**(HN)) can be characterized as follows:

DIII-21: x is an actual model of the architecture of the Hopfield network ($x \in$ **M**(HN)) iff there exist N, N_{in}, N_{out}, N_{hidden}, IR, θ, C, pred, succ, W, ext_n, net_n, act_n, out_n, State, E such that:

(1) $x=\langle N, N_{in}, N_{out}, N_{hidden}, IR, C, pred, succ, W, ext_n, net_n, act_n, out_n, State, E \rangle \in \mathbf{M_p}$(HN)

(2) for all $n \in N$ it holds:

(2.1) $N_{in} = pred$

(2.2) $N_{out} = succ$

(2.3) $N_{hidden} = succ \cap pred$

(3) $N = N_{in} \cup N_{out} \cup N_{hidden}$,

(4) $N_{hidden} \cap (N_{in} \cup N_{out}) = \varnothing$

(5) $N_{hidden} = \varnothing$

(6) $N_{in}=N_{out}=N$

(7) $C = N \times N - \{(n, n) | n \in N\}$.

(8) $n_1, n_0 \in N$, $n1 \neq n0$: $w_{n1,n0} = w_{n0,n1}$

(9) Input-net: $\forall n \in N$: $net_n = f_{net}^{(u)}(w^{\rightarrow}_n, in^{\rightarrow}_n) = w^{\rightarrow}_n\ in^{\rightarrow}_n = \Sigma_{n0 \in N-\{n1\}} w_{n0,n1}\ out_{n0}$.

(10) Activation

(10.1) $\forall n \in N$: $act_n = f_{act}^{(n)}(net_n, n)$

(i) $act_n = 1$, if $net_n \geq \theta_n$,or

(ii) $act_n = -1$, otherwise.

or

(10.2) $\forall n \in N$: $act_n = f_{act}^{(n)}(net_n, \theta)$

(i) $act_n = 1$, in case $net_n > \theta_n$, or

(ii) $act_n = -1$, in case $net_n < \theta_n$, or

(iii) $act_n = act_n$, in case $net_n = \theta_n$.

(11) output: $\forall n \in N: f_{out}^{(n)}(act_n) = act_n$

(12) $E = -1/2\ act^{\rightarrow T} W\ act^{\rightarrow} + \theta^{\rightarrow T}\ act^{\rightarrow}$

whereas: $act^{\rightarrow} = (act_{n1}, \ldots, act_{nm})^T$ is the activation state of the network.

W is the matrix of the weight of the Hopfield network.

$\theta^{\rightarrow} = (\theta_{n1}, ..., \theta_{nm})^T$ is the vector of the thresholds of the neurons.

The partial potential models of the Hopfield network can be derived from the potential models by omitting the **T**-theoretical concepts. In the Hopfield network, the terms net_n, act_n, out_n, *State,* and *E* are **T**-theoretical because they presuppose the Hopfield network itself. The net_n, act_n, out_n are **T**-theoretical concepts of the more general architecture of the neural network, and State and E are **T**-theoretical elements of the Hopfield network itself. The partial potential models of the Hopfield network ($\mathbf{M_{pp}}$(HN)) are characterized as follows:

DIII-22: y is a partial potential model of the Hopfield network ($y \in \mathbf{M_{pp}}$(HN)) iff there exists x such that:

(1) $x = \langle N, N_{in}, N_{out}, N_{hidden}, IR, \theta, C, pred, succ, W, ext_n, net_n, act_n, out_n, State, E \rangle \in \mathbf{M_p}$(HN)

(2) net_n, act_n, out_n, State, E are **T**-theoretical.

(3) $y = \langle N, N_{in}, N_{out}, N_{hidden}, IR, \theta, C, pred, succ, W, ext_n \rangle \in \mathbf{M_{pp}}$(HN)

3.3.3.3. The Theory-Net for the Network-Architecture

Until the current discussion, we just see as if the network-architecture of the artificial neural network has only two immediate specializations, i.e., the two layers feed-forward neural network and the Hopfield network. The fact is that the general network-architecture has several specializations in the form of feed-forward neural networks by adding the new statement: $N_{in} \cap N_{out} = \varnothing$ and recurrent neural networks by adding the new statement that there is a loop within the networks. The feed-forward neural networks have two specializations, namely the single-layer neural networks and the multi-layer neural networks. The specializations can be derived by adding some additional statements about the network and by applying several statements of the appropriate neuron models. The recurrent neural networks have many specializations, but here only the Hopfield network is discussed. We also get the Hopfield network by adding some additional statements and two theoretical terms about the network and by applying several statements of the appropriate neuron's models.

3.3.4. Building A Structuralist Model of the Delta Rule

In order to operate as expected, artificial neural networks need to learn. They can learn to minimize error according to specific rules/algorithms. The learning rule plays the role of guiding the artificial neural network to reach the optimal state of operation by adjustment of the synaptic weight or the threshold of the neurons. The goal is to reach the minimal error or cost in supervised learning and unsupervised learning or maximal payoff by reinforcement learning. Related to the topic of this dissertation, the delta rule, as the learning rule, will be applied for the two-layers feed-forward neural network because the connectionist simulation of cognitive dissonance discussed later implements this learning rule.

The basic idea of the learning rule used here is how to minimize error (normally called 'cost') by adjusting the weight of connectivity between neurons. If the neural network reaches the minimal cost, the learning step should be terminated. This condition is called convergence. Because of this goal, we need two basic functions, especially for feed-forward with the delta rule (and back-propagation), namely the cost-function and the gradient descent function.

The Widrow-Hoff Model of Learning Rule or the Delta Rule. According to the delta rule, the neural network should be fed with a set of inputs (*Inp*) and trained with expected outputs (*Out*) as a training set *L*. Normally the *Inp* and the *Out* are written in the form of matrices $\overrightarrow{Inp}$ and $\overrightarrow{Out}$. The connection's weight *W* and the threshold θ can be set randomly (Here we use the bias *B*, the bias is normally set =1 and its connection weight $w_0 \in IR$). With those inputs *INP*, the neural network will produce the actual output *OUTn*. In the first time of our network's computation, there will be a difference between the actual output (*OUTn*) and the value of the desired output ($\overrightarrow{Out}$). These value differences, called the error (*Error*), will be corrected by training our neural network according to the delta rule. We also assign a real number as a learning rate η. With the delta rule, we update the weight of every synaptic weight.

The general strategy of the Delta Rule is as follows: (1) We start with measuring the error in the output. It is defined by the difference between the actual output and the desired output according to the following formula:

$$\text{Error} = 1/n \, \Sigma \, (\text{out} - \text{outn})^2 \qquad \text{(the cost function)}$$

(2) The second step is to modify the weight to decrease the error of the network. Because the error is calculated for the whole pattern, the local error is not available. Therefore, we need to derive the error related to the activation of each output unit so that we can determine how the error will change according to each neuron's activation. This step can be done by calculating

the error locally as the difference between the desired and the actual output activation (I use the symbol δ for the local difference):

$\delta Error/\delta outn_n = (out - outn_n)$ (the derivation of Error by outn)

It tells us how far the output of each neuron input must be changed to minimize the error.

(3) Because we cannot change the input of the network (*Inp*), we must change the connection weight (*W*) in order to reduce the error. To do that, we can use the following chain rule:

$\delta Error/ \delta weight_n = \delta Error/ \delta outn_n . \delta outn_n / \delta weight_n$

(4) To evaluate the partial derivative of actual output *OUTn* related to the connection's weight (*W*), we have the linear activation rule:

$outn_n = \Sigma_i(w_{ni}\ inp_i)$ or $netn + b.w_0$ (activation-function of the perceptron)

For this partial derivative is:

$\delta outn_n / \delta weight_n = inp_i$

(5) The partial derivative of the error related to each weight (with negative sign) can be computed by multiplying the discrepancy by the input unit's activation.

$\delta Error/ \delta weight_n = -(out - outn_n).inp_i$

(6) The delta-rule multiplies this by the learning rate η

$\Delta w_i = -\eta\ (out - outn_n).inp_i$

Alternatively, for the network:

$\Delta w_i = -\Sigma_i\ \eta\ (out - outn_{ni}).inp_i$

Note: If we set the learning rate too small, the learning process will be very slow. However, if we set the learning rate too big, then the neural network will not reach the minimal cost (convergent state) because it takes a too big step in the gradient descent.

(7) The last step is to update the connection weight by the following rule:

$w_i^{(neu)} = w_i^{(alt)} + \Delta w_i$.

Borgelt et al. give the Convergence-statement for Delta rule as follows (Borgelt, C. et al., 2003, p. 27):

"*Sei* $L=\{(x_1^{\rightarrow}, o_1),..., (x_m^{\rightarrow}, o_m)\}$ *eine Menge von Trainingsbeispielen, jeweils bestehend aus einem Eingabevektor* $x_i^{\rightarrow} \in IR^n$ *und der zu diesem Eingabevektor gewünschten Ausgabe* $o_i \in \{0,1\}$. *Weiter sei* $L_0 = \{(x^{\rightarrow}, o) \in L | o = 0\}$ *und* $L_1 = \{(x^{\rightarrow}, o) \in L | o = 1\}$. *Wenn* L_0 *und* L_1 *linear separabel sind, d.h., wenn* $w^{\rightarrow} \in IR^n$ *und* $\theta \in IR$ *existieren, so dass*

$$\forall (x^{\rightarrow}, 0) \in L_0: \quad w^{\rightarrow} x^{\rightarrow} < \theta \text{ and}$$

$$\forall (x^{\rightarrow}, 1) \in L_1: \quad w^{\rightarrow} x^{\rightarrow} \geq \theta,$$

Dann terminieren [der Trainingsalgorithmus]

[In English: Let $L=\{(x_1^{\rightarrow}, o_1),..., (x_m^{\rightarrow}, o_m)\}$ be a set of training examples, each one consisting of an input vector $x_i^{\rightarrow} \in IR^n$ and for this vector input there is a desired output $o_i \in \{0,1\}$. Further let let $L_0 = \{(x^{\rightarrow}, o) \in L | o = 0\}$ and $L_1 = \{(x^{\rightarrow}, o) \in L | o = 1\}$. If L_0 and L_1 are linearly separable, i.e., if there exist $w^{\rightarrow} \in IR^n$ and $\theta \in IR$, such that:

$$\forall (x^{\rightarrow}, 0) \in L_0: \quad w^{\rightarrow} x^{\rightarrow} < \theta \text{ and}$$

$$\forall (x^{\rightarrow}, 1) \in L_1: \quad w^{\rightarrow} x^{\rightarrow} \geq \theta,$$

Then [the training algorithm for the delta rule] is terminated.]."

This dissertation uses the symbol *B* for bias instead of the symbol θ.

The Structuralist Model of the Delta Rule. To build a structuralist model for the delta rule we characterize the potential models, the actual models, and the partial potential models. The potential models of the delta rule ($\mathbf{M_p}$(DR)) can be formulated as follows:

DIII-23: x is a potential model of the delta rule ($x \in \mathbf{M_p}$(DR)) iff there exist N, IR, Inp, Out, C, B, W, OUTn, η, Error so that:

(1) $x = \langle N, IR, Inp, Out, C, L, B, W, OUTn, \eta, Error \rangle \in \mathbf{M_p}$(DR)

(2) N is a finite non-empty set of neurons.

(3) $Inp \subset IR$ (Input)

(4) $INP^{\rightarrow}$ is a set of the input vector.

(5) $Out \subset IR$ (desired Output)

(6) $OUT^{\rightarrow}$ is a set of an output-vector.

(7) $C \subseteq N \times N$ is a finite non-empty set of directed connections between neurons.

(8) $L \subseteq Inp \times Out$ (a finite non-empty set of training examples)

(9) $B := N \rightarrow IR$ (Bias)

(10) $W := C \rightarrow IR$ (weight)

(11) $Outn \in IR$ (actual output, if the neural network is fed with input Inp)

(12) $\eta \in IR$ (learning rate)

(13) $Error := Out \times OUTn \rightarrow IR^2$ (The network's error is a mapping into a two-dimensional Cartesian coordinate system)

The actual models of the delta rule (**M**(DR)) consist of the following law statements or law-like statements: (1) The network's error can be calculated by dividing the sums of the square of the discrepancy between actual output and expected output by two. This method to calculate the network's error is known as the mean square error (MSE) method. To correct each neuron's error, we need to derive this network's error for each neuron. (2) The learning-rule is as follows: (a) For all neurons' bias, the update is $B^{(old)} + \Delta B$ with $\Delta B = -\eta(Out - OUTN)$. (b) For all neurons' connection-weight the update is $w_i^{(new)} = w_i^{(old)} + \Delta w_i$ with $\Delta w_i = \eta(Out - OUTN)\, Inp_i$ and η is the learning rate. (3) The Convergence-statement for the delta rule is as follows: given a set of training-sample L containing pairs of input and desired output $L=\{(Inp_1^{\rightarrow}, Out_1),..., (Inp_m^{\rightarrow}, Out_m)\}$, the set L contains $L_0 = \{(Inp^{\rightarrow}, Out) \in L | Out = 0\}$ and $L_1 = \{(Inp^{\rightarrow}, Out) \in L | Out = 1\}$. If L_0 and L_1 are

linearly separable, and if $w^{\rightarrow} \in IR^n$ and B=IR exist, then for all L_0 the net-input is < B and for all L_1 the network input is ≥ B.

The actual models for the delta rule (**M**(DR)) can be defined as follows:

DIII-24: x is an actual model of the delta rule ($x \in$ **M**(DR)) iff there exist N, IR, Inp, Out, C, L, B, W, OUTN, η, Error such that:

(1) $x = \langle N, IR, Inp, Out, C, L, B, W, OUTn, \eta, Error \rangle \in \mathbf{M_p}(DR)$

(2) $\forall i \in \{1, ..., n\}$: $Error = \frac{1}{2} \Sigma_i (Out - OUTN)^2$ and the derivation for each neuron's activation: (Out – OUTN)

(3) $\forall b_i \in B$, i=1, ..., n: $b_i^{(new)} = b_i^{(old)} + \Delta b_i$ with $\Delta b_i = -\eta(Out - OUTN)$.

(4) $\forall w_i \in W$, i=1, ..., n: $w_i^{(new)} = w_i^{(old)} + \Delta w_i$ with $\Delta w_i = \eta(Out - OUTN)\ Inp_i$

(5) Convergence-statement:

Supposed $L=\{(Inp_1^{\rightarrow}, Out_1),..., (Inp_m^{\rightarrow}, Out_m)\}$ is a set of training-sample with

$L_0 = \{(Inp^{\rightarrow}, Out) \in L | Out = 0\}$ and $L_1 = \{(Inp^{\rightarrow}, Out) \in L | Out = 1\}$.

If L_0 and L_1 are linearly separable and if $w^{\rightarrow} \in IR^n$ and B =IR exist, then

$\forall (Inp^{\rightarrow}, 0) \in L_0$: $w^{\rightarrow} Inp^{\rightarrow} < B$ and

$\forall (Inp^{\rightarrow}, 1) \in L_1$: $w^{\rightarrow} Inp^{\rightarrow} \geq B$.

The partial potential models of the delta rule can be characterized by omitting the **T**-theoretical elements from the potential model $\mathbf{M_p}$(DR), which are the learning-rate (η) and the error (*Error*) because both terms presuppose the delta rule itself. The learning rate (η) determines the learning's speed. The *Error* is here understood as the difference between actual outputs and the desired outputs in the output-layer. Therefore, the partial potential models of the delta rule ($\mathbf{M_{pp}}$(DR)) can be defined as follows:

DIII-25: y is a partial potential Model of the delta-rule ($y \in \mathbf{M_{pp}}$(DR)) iff there exist x such that:

(1) $x = \langle N, C, Inp, Out, L, B, W, OUTN, \eta, Error \rangle \in \mathbf{M_p}$(DR).

(2) η and Error are **T**-theoretical.

(3) $y = \langle N, C, Inp, Out, L, B, W, OUTN \rangle \in \mathbf{M_{pp}}$(DR).

Chapter 4
Some Preliminary Work for Building the Structuralist Models of Intertheoretical Connections in Some Cases in Cognitive Science

Before starting with modeling intertheoretical connections between several theories in cognitive science, some preparations in this chapter should be made. The preparations involve some adjustments to the structuralist theory of science and an overview of cognitive science relevant to this dissertation.

4.1. Some Adjustments in the Structuralist Theory of Science

Though the standard version of the structuralist metatheory of science in BMS is a powerful tool to represent scientific theories and their intertheoretical connections formally, two improvements are needed to deliver a better analysis of intertheoretical connections in interdisciplinary fields. The first improvement is the notion of echelon partial substructure developed by Moulines in his paper *Intertheoretical Relations and the Dynamics of Science*, 2014. And the second is a revision that I propose for the definition of a specialization.

4.1.1. The Notion of Echelon Partial Substructure

In 2014, Moulines gave a formal definition of echelon partial substructure as a preparation for giving a formal structuralist account for four types of theoretical changes from the diachronic point of view, i.e., in the development of scientific theories. The notion of echelon partial structure is intuitively as follows: Given a set-theoretic operation Θ, which consist in successively applying a finite number of times the operations of power-set

construction and cartesian product to some given sets. The operation Θ always begins with power-set construction. S is an echelon partial substructure of S* iff for all S_i components of S, there exist some S^*_k component(s) of S* where S_i is the range of operation Θ applied to S^*_k. The complete definition of echelon partial substructure can be seen in Moulines, 2014, p. 1512. This definition is beneficial to characterize subsets of potential models ($\mathbf{M_p}$) that contents components of $\mathbf{M_p}$, which some certain intertheoretical connection are applied to. Respectively this definition can also be applied to $\mathbf{M_{pp}}$ for characterizing (local) empirical claims of the intertheoretical connection.

With the definition of echelon partial substructure, Moulines can distinguish in a precise manner between concepts that are connected to one another in a diachronic intertheoretical relation and concepts which are not. The notion of echelon partial substructure will also be applied here for the same reason to some cases of synchronic intertheoretical relations. To build several models of intertheoretical relations precisely, we need not only all definitions of intertheoretical connection and its varieties of Chapter 2, but also the definition of echelon partial substructure. The reason for this is that in many cases, there are some unconnected concepts and some other concepts in the potential models, which connect to specific concepts in another theory-element through intertheoretical relations – These concepts must be distinguishable. Such cases can also be seen in most cases of intertheoretical reduction in Chapters 5–7. They are examples of cases of partial reduction, of which only several concepts in the potential models ($\mathbf{M_p}$) of a theory element **T**, as a higher-level theory, can be reduced by concepts in the potential models of another theory element **T***, a lower-level theory. In the next three chapters, we will see that not all concepts in the potential models of Festinger's theory $\mathbf{M_p}$(DissB) (or in the potential models of forced-compliance dissonance $\mathbf{M_p}$(DissF)) can be reduced by the concepts of the potential models of the Hawkins-Kandel Computational Neurobiological

Theory $\mathbf{M_p}$(CNT), the Hopfield network $\mathbf{M_p}$(HN), or the two-layer feed-forward neural network $\mathbf{M_p}$(RP+2L-FFN+DL). The notion of echelon partial substructure will be implemented in order to characterize which concepts in $\mathbf{M_p}$(DissB) or $\mathbf{M_p}$(DissF) are being reduced by the concepts of $\mathbf{M_p}$(CNT), $\mathbf{M_p}$(HN), or $\mathbf{M_p}$(RP+2L-FFN+DL).

Therefore, the procedure for our modeling and analysis will be as follows: Generally, a definition of special types of intertheoretical connection is applied to create a model of intertheoretical connections for our selected cases. However, the definition of both determining and entailment links are used to get a more detailed analysis of intertheoretical relations. By using the definition of both basic types of links, we can identify all connected concepts and build an echelon partial substructure of the potential models of the connected theories. Moreover, we will analyze how those intertheoretical connections work and connect the terms of those theories with respect to the **T**-non-theoretical level of the connected theories.

4.1.2. A Revision of the Definition of Specialization

The second improvement is a revision of the definition of intertheoretical specialization DIV-1 in the BMS, p. 170. As mentioned in Chapter 3, Rainer Westermann had already built a theory-net of the Festinger theory of cognitive dissonance with its four specializations. Let us now look at one of its specializations, called the forced compliance dissonance. In this case, the specialization relation is built not only by adding some new law statements or law-like statements to the actual models of the Festinger theory of cognitive dissonance **M**(DissB), but also by modifying its potential models $\mathbf{M_p}$(DissB) (see DIII-1 and DIII-2). The modification in the potential models is made by adding several restrictions, which make the extension of the potential models of forced compliance dissonance $\mathbf{M_p}$(DissF) (see DIII-4) narrower than the $\mathbf{M_p}$(DissB).

Similar cases also occur in three other specializations of the Festinger theory of cognitive dissonance, namely post-decision dissonance (DissD), information exposure dissonance (DissI), and social disagreement dissonance (DissS). In these three cases, additional restrictions are not only added to the actual model, but also to the potential model. $\mathbf{M_p}$(DissD), $\mathbf{M_p}$(DissF), $\mathbf{M_p}$(DissI), and $\mathbf{M_p}$(DissS) are no longer equal to $\mathbf{M_p}$(DissB), but they become the subsets of $\mathbf{M_p}$(DissB). Therefore, $\mathbf{M_{pp}}$(DissD), $\mathbf{M_{pp}}$(DissF) (see DIII-6), $\mathbf{M_{pp}}$(DissI), and $\mathbf{M_{pp}}$(DissS) are also the subsets of $\mathbf{M_{pp}}$(DissB) (see DIII-3). From these cases, the author finds that the definition of specialization D IV-1 in BMS, p. 170, is too strong – especially with respect to the condition DIV-1 (1) $\mathbf{M_p' = M_p}$ and $\mathbf{M_{pp}' = M_{pp}}$. This definition does not provide a possibility to modify the $\mathbf{M_p}$ by adding some more restrictions, which will be able to produce a type of specialization as well. Therefore, this dissertation suggests a modification to the definition of specialization as follows: (this definition will be used for the rest of this dissertation).

D IV-1: If $\mathbf{T = \langle M_p, M, M_{pp}, GC, GL, I\rangle}$ and $\mathbf{T' = \langle M_p', M', M_{pp}', GC', GL', I'\rangle}$ are idealized theory-elements, then **T'** is an idealized specialization of **T** (abbreviated as $\mathbf{T'\sigma T}$) iff:

(1) $\mathbf{M_p' \subseteq M_p}$ and $\mathbf{M_{pp}' \subseteq M_{pp}}$,

(2) $\mathbf{M' \subseteq M}$, $\mathbf{GC' \subseteq GC}$, $\mathbf{GL' \subseteq GL}$ and $\mathbf{I' \subseteq I}$,

This modification still retains the three characteristics of the specialization relation in Theorem IV-1 in BMS page 170, but diminishes a possible tension between DIV-1 (1) $\mathbf{M_p' = M_p}$ and $\mathbf{M_{pp}' = M_{pp}}$ and TIV-18 (b) $\mathbf{Cn(K') \subseteq Cn(K)}$ of BMS. Also, it will be able to solve, or at least reduce, tension with the definition of intertheoretical connections as a bridge between theories in Moulines and Polanski, 1996, p. 222. In the new definition of specialization, there must not appear two statements that seemingly contradict each other: On the one hand $\mathbf{M_p' = M_p}$ and $\mathbf{M_{pp}' = M_{pp}}$ (BMS, 1987, p. 170)

and, on the other hand, the definition of an intertheoretical connection $\exists i,j\ (1 \leq i,j \leq n \wedge \mathbf{M_p}^i \neq \mathbf{M_p}^j)$ (Def 6(2) in Moulines and Polanski, 1996, p. 222).

4.2. An Overview of Cognitive Science Related to This Project

Although cognitive science is a relatively young interdisciplinary field, it has already been a very fruitful scientific field. There are many research programs in this field that are very helpful to provide us with a more comprehensive and in-depth explanation of phenomena of the mind, or that inspire us to develop many applications in the form of technologies and techniques (such as artificial neural networks for face or speech recognition, predictions for the stock exchange, and others). Cognitive science is an interdisciplinary field that studies phenomena of cognition – not only limited to human cognition but including animal cognition as well. There are at least six scientific fields that constitute cognitive science. They are philosophy (of mind), psychology, neuroscience, linguistics, artificial intelligence, and anthropology – or other social sciences. (Figure 4.1)

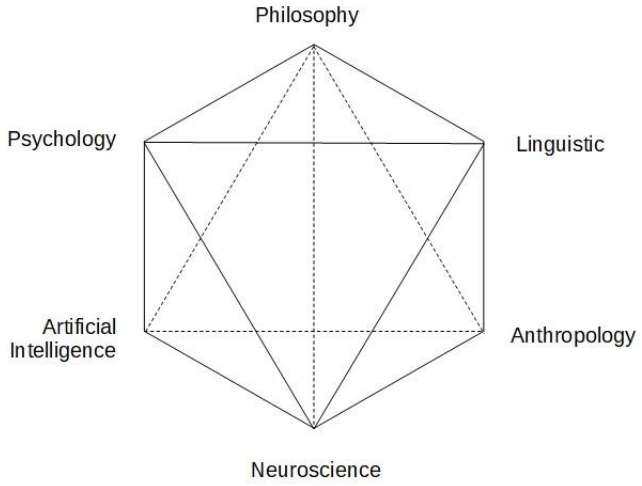


Figure 4.1. The fields in cognitive science according to Keyser et al. 1978; Solid lines indicate near or strong connections and dashed lines indicate far or weaker connections. (Source: Stephan, Achim, and Walter, Sven, 2013, p.3)

In this dissertation, our modeling of intertheoretical connections in cognitive science will cover only three areas, namely psychology, neuroscience, and artificial intelligence. This modeling is based on the following description. Cognition comprises very complex phenomena that have been difficult to comprehend completely ever since ancient times. In the philosophy of mind, there are many theories that try to explain the mind (or cognition) and its relationship with the body. An interest in the mind and behavior can be found in the ancient civilizations. At that time, psychology was a part of philosophy. Psychology only started to be an independent field in 1879 when Wilhelm Wundt, who called himself as a psychologist, built the first laboratory for psychological research in Leipzig. Since then, many approaches and schools in psychology have been founded (especially in cognitive psychology) to explain the phenomena of mind and their aspects, such as behaviorism, psychoanalysis, and many more.

In cognitive science and modern philosophy of mind, all these schools are called folk psychology or commonsense psychology. The term *folk psychology* means "(1) commonsense psychology that explains human behavior in terms of beliefs, desires, intentions, expectations, preferences, hopes, fears, etc.; (2) an interpretation of such everyday explanations as part of a folk theory, comprising a network of generalizations employing concepts like belief, desire, and so on" (Baker, 1999, p. 319). In cognitive science, this folk psychology is occasionally seen as an anti-scientific view of our self-understanding and, therefore, replaceable by other approaches related to neuroscience and artificial intelligence. Several cognitive scientists, such as Stephen P. Stich, Paul M. Churchland, and Patricia R. Churchland, who try to combine, or respectively reduce psychology to neuroscience, call themselves eliminative materialists. On the other hand, many philosophers admit a kind of reduction relation between psychological processes and brain processes but do not demand replacement of folk psychology theories, such as Jaegwon Kim, Terence Horgan, James Woodward, Daniel Dennett, among others.

A second important discipline in the study of the mind is neuroscience, especially cognitive neuroscience. Neuroscience is the study of the brain and its neural networks explaining how our cognitive process takes place in our brains. In cognitive neuroscience, the study of the cognitive process is done in two ways. First, to discover the functionality of brain parts or regions in the cognitive process, neuroscientists observe the correlation between disturbances of cognitive capacities (e. g. aphasia) or personality-changes and damage to certain parts of the brain (e. g. lesions). From such observations, neuroscientists can identify the functionality of certain parts of the brain in cognitive processes. The second way is to do experiments – via models and observations – of how the neurons and their network produce certain aspects of cognition, of how the metabolism of a brain works and what its influence is in cognitive processes. The research is carried out not only on human brains but also on animal brains.

There are a huge number of neurons and their connections in a brain; for example, the human brain contains around 10^{11} neurons and 10^{14} synapses, and even now, it is impossible to identify all existing connections (connectome), especially in vivo. Sebastian Seung writes poetically:

> "No road, no trail can penetrate this forest. The long and delicate branches of its trees lie everywhere, choking space with their exuberant growth. No sunbeam can fly a path tortuous enough to navigate the narrow spaces between these entangled branches. All the trees of this dark forest grew from 100 billion seeds planted together. And, all in one day, every tree is destined to die. This forest is majestic, but also comic and even tragic. It is all of these things. Indeed, sometimes I think it is everything. Every novel and every symphony, every cruel murder and every act of mercy, every love affair and every quarrel, every joke and every sorrow— all these things come from the forest. You may be surprised to hear that it fits in a container less than one foot in diameter. And that there are seven billion on this earth. You happen to be the

> caretaker of one, the forest that lives inside your skull. The trees of which I speak are those special cells called neurons. The mission of neuroscience is to explore their enchanted branches— to tame the jungle of the mind" (Seung, 2012).

To study how a brain works and produces cognition, cognitive scientists assume that a brain is a black box. They use artificial intelligence as an aid to model and simulate a cognitive process. One of the approaches in artificial intelligence that is often implemented is called artificial neural networks. It is so-called because it is strongly inspired by the concept of a neuron.

The intertheoretical relations of these theories will be formally modeled according to the following interdisciplinary relations. Under the topic of interdisciplinary relation between psychology and neuroscience, the intertheoretical connections between the theory of forced compliance dissonance in psychology and the Hawkins-Kandel computational neuroscientific theory (CNT) in neuroscience will be modeled based on the result of research by Vincent van Veen, et al. The second interdisciplinary relation is between psychology and artificial intelligence. We will model the intertheoretical connections of two simulations of cognitive dissonance, namely the consonance and the connectionist models. The consonance model implements the Hopfield network to simulate the phenomena of dissonance reduction. Therefore, we are going to model the intertheoretical connections between the Festinger (general) theory of cognitive dissonance and the Hopfield network. For the connectionist model, which uses the two-layers feed-forward neural network with the delta rule, we are going to model the intertheoretical connection between the Rosenblatt perceptron, the two layers feed-forward neural network and the delta rule on one side and the forced compliance dissonance theory on the other side. And the last interdisciplinary relation is between neuroscience and artificial intelligence, namely between the McCulloch-Pitts neuron and the Rosenblatt perceptron. These concrete cases of intertheoretical relations can be seen in Figure 4.2.

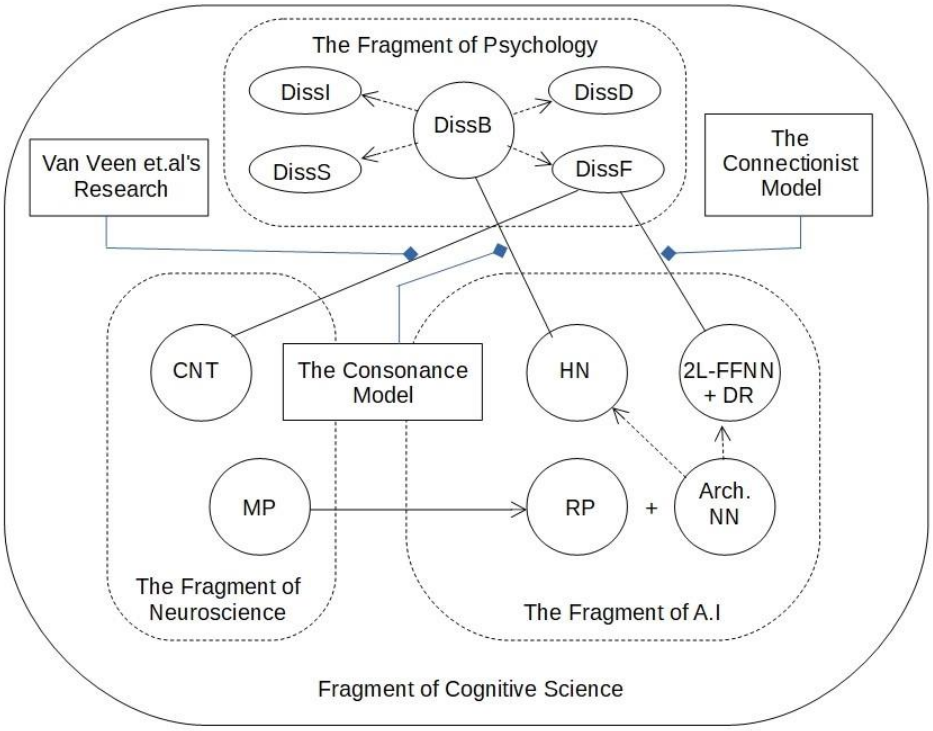


Figure 4.2. The map of the intertheoretical relations, that will be discussed in this dissertation; CD = the theory of cognitive dissonance (DissB model); DissD = the post decision dissonance theory; DissF= the forced compliance dissonance theory; DissI = the dissonance & information exposure theory; DissS= the social disagreement dissonance theory; CNT = the computational neuroscientific theory; MP= the McCulloch-Pitts neuron; RP= the Rosenblatt perceptron; Arch.NN= the architecture of neural network model; HN=the Hopfield network; FFNN+DR= the feed-forward neural network model and the delta rule; continuous lines = interdisciplinary relations/connections; continuous arrows = the specialization Relations in a theory-net. The researches modeled are VanVeen et.al's (DissF & CNT), the consonance model (DissB & HN), and the connectionist model.

Chapter 5

The Structuralist Model of Intertheoretical Connections between the Festinger Theory of Cognitive Dissonance and the Hawkins-Kandel Computational Neuroscientific Theory in the Process of Dissonance Reduction in the *Dorsal Anterior Cingulate Cortex (dACC)*

5.1. Van Veen's Research Program and the Connection between Psychology and Neuroscience

Many scientists and philosophers believe that there are connections between human psychological phenomena and how the human brain works. They develop some relations between psychological theories and theories about how the brain works – intertheoretical relations in these matters are known as intertheoretical reduction. Many scientists attempted to show how experiments can confirm such intertheoretical relations. The first intertheoretical relation that will be modeled and analyzed is the relation between Festinger's theory of cognitive dissonance and Hawkins-Kandel's computational neuroscientific theory (CNT) by putting in context of a research conducted by Vincent van Veen et al. (2009). This analysis aims to show how far such a (widely) believed intertheoretical relation between two theories from different disciplines could be confirmed by research about the phenomena explained by it. In our case, both theories are from psychology and neuroscience. Van Veen et al. themselves had no intention to prove the intertheoretical reduction between Festinger's theory of cognitive dissonance and the Hawkins-Kandel computational neuroscientific theory (CNT). Their research aims to show that a specific area of *cerebral cortex* in the human brain plays a vital role during a phase of cognitive dissonance.

In 2009 Van Veen et al. investigated a specific area of the cerebral cortex called *dorsal Anterior Cingulate Cortex (dACC)* or Broadman's area 32. The Brodmann areas are areas of the cerebral cortex in human or other primate brains defined by the neurons' cytostructure. The division and numbering of the Broadman areas were initially made by the German anatomist Korbinian Brodmann in 1909. The division of the Brodmann areas has been refined and renamed for more than a century and remains widely used as a reference of the cytoarchitectural organization of the human cortex.

The *dACC* is a part of the *Anterior Cingulate Cortex (ACC)* that deals with cognitive processes, such as reward anticipation and decision making. The other area of the *ACC* is called *ventral Anterior Cingulate Cortex (vACC)*, which deals with emotional processes. The *dACC* is connected to the prefrontal cortex, the parietal cortex, the motor systems, and the frontal eye fields. This position gives it a central role in processing top-down and bottom-up stimuli and assigning appropriate control to other areas of the brain.

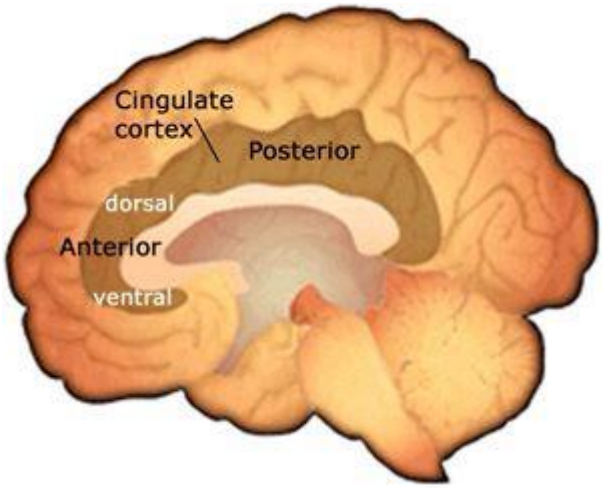


Figure 5.1. The *dorsal Anterior Cingulate Cortex* (Source: quizlet.com)

Because of its position, Van Veen and others proposed that the *dACC* has a function to detect conflicts between prior attitudes and counter-

attitudinal behavior in cognitive dissonance. “One of the *dACC*’s functions in cognition is to detect the conflict between active, but incompatible, streams of information processing, such as between the color and the meaning of a word in the Stroop task. *dACC* activation is consistently related to the amount of conflict occurring in such tasks. … We hypothesized that the *dACC*’s conflict monitoring function might generalize from detecting conflict in simple speeded-response tasks to detecting the conflict between prior attitudes and counter-attitudinal behavior in cognitive dissonance” (van Veen et al., 2009, p. 1469). They also said that these functions could be simulated by implementing the Hopfield network: “Computational simulations of conflict in simple speeded response tasks have measured conflict as Hopfield's energy and have shown that *dACC* activation in such tasks can be well modeled by this measure” (van Veen et al., 2009, p. 1469).

Van Veen et al. reported their observations using fMRI on the participants of an experiment of cognitive dissonance by applying the induced compliance procedure. The research was conducted as follows (van Veen et al., 2009, pp. 1469–1470): In the first step, participants performed a rather long (45 min) and tedious task in an uncomfortable environment, namely in a magnetic resonance scanner. While participants performed this task, they were randomly assigned to one of two groups: the dissonance group or the control group.

In the second step, the participants had to respond to sentences presented on a screen with their left or right ring, middle, or index finger, that represented 6 points on the Likert scale (1 = left ring finger, agree entirely; 6 = right ring finger, disagree entirely). There were two types of sentences: target sentences, which consist of attitudes toward the scanner, and neutral sentences, composed of other topics or tasks. Participants in the control group were told to respond to the target sentences as if they were enjoying the scanner and the task, regardless of their actual feeling about the experience. They were informed that they would receive additional money for each

sentence, to which they responded as if they were enjoying the scanner and the task. However, they were told to respond honestly to the other sentences, namely the neutral sentences.

Participants in the dissonance group were given instructions about how to respond to the stimuli. They were told that a patient had been scheduled to be scanned after them and was to perform a similar task in the scanner. The participants were informed that this patient was now in the control room, watching the experimental control computer screen, and was very nervous and uncomfortable about his upcoming scanning session. The participants were informed that several of the sentences were about their attitudes toward the scanner and the task. They were asked if they would be willing to respond as though they enjoyed being in the scanner and doing the task, regardless of their real feelings about the experience. They were made to believe that the patient in the control room could see the responses on the screen. This situation was set to create a counter-attitudinal argument.

In the third step, participants were led into a private waiting room after the scanning was completed. There they were asked to fill out a set of forms, which also contained the target sentences. This time they were asked to respond according to their actual feeling about their experience in the scanner. A composite score was calculated for the participants' enjoyment of the scanner and task. After completing the forms, the participants were carefully interviewed; participants who admitted having doubts about the validity of the cover story were not included in the analysis.

The result was as follows: "We found that participants in the dissonance group changed their attitudes more than participants in the control group following counter-attitudinal behavior. Furthermore, *dACC* and *anterior insula* activation during counter-attitudinal behavior predicted the participants' final attitude in the dissonance group, but not in the control group. In the dissonance group, these partial correlations were significant for the bilateral *dACC* and bilateral anterior insula regions (partial r range = 0.60–

0.68, all P < 0.01); for those regions, these correlations were not significant in the control group (partial r range = –0.33–0.11, all P > 0.1). ANCOVA analysis verified that for the bilateral *dACC* and left *Anterior Insula*. These correlations were greater in the dissonance group than in the control group (F1,35 range = 4.10–9.43, all P < 0.05). For the right anterior insula, the ANCOVA was marginally significant (F1,35 = 3.186, P = 0.083)" (van Veen et. al, 2009, p. 1472).

In the discussion part of their article, van Veen, et al. (2009) write: "These findings are consistent with a number of prior observations. Both cognitive dissonance and *dACC* and anterior insula activation have been associated with negative affect and autonomic arousal. These regions might, therefore, be responsible for representing or triggering the negative affect and related autonomic arousal associated with the dissonance. … Our data expand on those findings, indicating that *dACC* activity during the counter-attitudinal argument, which is similar to lying, predicts subsequent attitude change, but only when counter-attitudinal behavior conflicts with other cognitions. … In short, our results are consistent with theories of cognitive dissonance that emphasize the conflict between different cognitions, such as the original theory. In particular, our results are consistent with the action-based model of cognitive dissonance, which posits that conflict between cognitions evokes an aversive state because it potentially interferes with unconflicted, effective, goal-driven action" (van Veen et al., 2009, p. 1472).

Based on van Veen et al.'s experiment, the Festinger theory of cognitive dissonance will be connected to the Hawkins-Kandel computational neuroscientific theory (CNT) according to the following principles. (1) Cognitions in the dissonance theory are defined as patterns of activation value or patterns of synaptic weight values that regulate activation values in all, except the input neurons of the network (Bickle, 1998, p. 191). (2) The network of neurons discussed here will be limited to the network of neurons

in *dACC*. (3) The research conducted by van Veen shows that cognitive dissonance is related to high activation of the *dACC* area and *anterior insula*.

5.2. Building A Structuralist Model of the Intertheoretical Connection between the Festinger Theory of Cognitive Dissonance and the Hawkins-Kandel Computational Neuroscientific Theory (CNT)

To build a model of the intertheoretical relation between the Festinger theory of cognitive dissonance and the computational neuroscientific theory, we follow the idea that cognition at the psychological level can be characterized as a pattern of activation values and synaptic weight values that regulate activation values in all neurons of the network. Therefore, the term *Cognition* in the theory of cognitive dissonance is related to the relation between activation values (*AV*) and connection weights (*CW*) of the computational neuroscientific theory. Because both theories assume that both cognition and this activation and connection weights occur at certain discrete point(s) of time, the terms time (*T*) of both theories are connected to each other.

To make this model more concrete, we are going to refer to van Veen et al.'s research about the connections between a special case of cognitive dissonance, namely forced compliance dissonance, and the neural activities in a particular brain region called the *dorsal Anterior Cingulate Cortex (dACC)* (for the detail experiment see Chapter 4). This brain area shows different activities between the event(s) of cognitive dissonance and event(s) of consonance. The experiment showed high activity of *dACC* and *Anterior Insula* during counter-attitudinal behavior. The terms *Disscog* and *Conscog* in the theory of forced compliance dissonance are connected to the high neural activities or the low neural activities of *dACC*.

Based on the case observed by van Veen et al., we have to build a model of the intertheoretical relation between a specialization of cognitive dissonance (DissB), called forced compliance dissonance (DissF), and the

computational neuroscientific theory (CNT). The specialization for the forced compliance dissonance (DissF) has several special terms, such as *Forcecom*, *attidiff*, *imp*, and *reward*. The *Forcecom* is a subset of cognition. Therefore, it must be connected to the relation between the set of activation values (*AV*) and the set of connection weights (*CW*). However, we have several uncertainties about connecting the *attidiff*, *imp*, and *reward* with terms in CNT. This fact is understandable because we cannot give such empirical evidence[1].

Because only several terms in the potential models of forced compliance dissonance ($\mathbf{M_p}$(DissF)) are connected with the potential models of the computational neuroscientific theory ($\mathbf{M_p}$(CNT)), the definition of determining link is used to identify the intertheoretical relation between the terms of both models. By identifying all determining links between them, we can determine the echelon partial substructure from the potential model of both theory-elements. The intertheoretical connections between both echelon subsets are now a kind of entailment link (see the definition of entailment link in Moulines & Polanski, 1996, p. 223). Based on these entailment links, we can apply the definition of interpreting links to connect the respective elements in the partial potential models of the forced compliance dissonance $\mathbf{M_{pp}}$(DissF). In this way, we determine the local intended applications of the intertheoretical connections.

The intertheoretical connections between the forced compliance dissonance and the computational neuroscientific theory for neurons in the *dorsal Anterior Cingulate Cortex (dACC)* can be formally characterized as follows: Since in this case, the computational neuroscientific theory is applied to explained the behavior of neurons in the *dorsal Anterior Cingulate Cortex (dACC)*, I call the theory-element of CNT in this case "CNT on *dACC*."

[1] with our technology today. In some simulations of dissonance reductions implementing artificial neural networks, the activation value of the artificial neurons is usually connected with the presence or absence of cognitions.

D V-1: If **T**(DissF) = ⟨**M**$_{\mathbf{p}}$(DissF), **M**(DissF), **M**$_{\mathbf{pp}}$(DissF), **I**(DissF)⟩ and **T**(CNT on *dACC*) = ⟨**M**$_{\mathbf{p}}$(CNT on *dACC*), **M**(CNT on *dACC*), **M**$_{\mathbf{pp}}$(CNT on *dACC*), **I**(CNT on *dACC*)⟩ then there exist Λ as a set of determining links λ_1, λ_2, λ_3, λ_4, λ_5 between T(DissF) and T(CNT on *dACC*) iff there exist x_1, x_2 such that:

(1) x_1 = ⟨Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, attidiff, imp, reward⟩ ∈ **M**$_{\mathbf{p}}$(DissF)
(Let x_1 be a potential model of the forced compliance dissonance (**M**$_{\mathbf{p}}$(DissF))

(2) x_2 = ⟨N, Act, T, IN, IR, AV, O, I, CW, Cause⟩ ∈ **M**$_{\mathbf{p}}$(CNT on *dACC*)
(Let x_2 be a potential model of the computational neuroscientific theory (**M**$_{\mathbf{p}}$(CNT on *dACC*))

(3) Time λ_1 T
(Let λ_1 be the first determining link, which connects the set *Time* in **M**$_{\mathbf{p}}$(DissF) to the set *T* in **M**$_{\mathbf{p}}$(CNT on *dACC*). λ_1 is bijective)

(4) Cognition λ_2 (AV × CW)
(Let λ_2 be the second determining link, which connects the set *Cognition* in **M**$_{\mathbf{p}}$(DissF) to the relation between the set of activation value *AV* and the set of connection weight *CW* in **M**$_{\mathbf{p}}$(CNT on *dACC*). The relation between *AV* and *CW* is defined according to DIII-17(2). It is also relevant for defining the λ_3)

(5) Forcecom λ_3 (AV × CW), where $\lambda_3 \subseteq \lambda_2$
(Let λ_3 be the third determining link, which connects the set *Forcecom* in **M**$_{\mathbf{p}}$(DissF) to the relation between the set of activation value *AV* and the set of connection weight *CW* in **M**$_{\mathbf{p}}$(CNT on *dACC*))

(6) Disscog λ_4 (CW × CW), where for all neurons in the network at a certain period of time Σ_i^m ||CW(t_i) - CW(t_{i+1})|| is big.

(Let λ_4 be the fourth determining link, which connects the set *Disscog* in $\mathbf{M_p}$(DissF) to the relation (difference) between connection weights *CW* at t_i and connection weights *CW* at t_{i+1} in $\mathbf{M_p}$(CNT). It is *Disscog* if for all neurons in the network at a certain period of time Σ_i^m $||CW(t_i) - CW(t_{i+1})||$ is big)

(7) Conscog λ_5 (CW × CW), where for all neurons in the network at a certain period of time Σ_i^m $||CW(t_i) - CW(t_{i+1})||$ is small.

(Let λ_5 be the fifth determining link, which connects the set *Conscog* in $\mathbf{M_p}$(DissF) to the relation (difference) between connection weights *CW* at t_i and connection weights *CW* at t_{i+1} in $\mathbf{M_p}$(CNT on *dACC*). It is *Conscog* if for all neurons in the network at a certain period of time Σ_i^m $||CW(t_i) - CW(t_{i+1})||$ is small)

(8) $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$

(Let L is a set of determining links between T(DissF) and T(CNT on *dACC*))

With the determining links defined above, the echelon partial substructure of both theory-elements, which represent the intertheoretically connected concepts between both theories, can be characterized. Both echelon partial substructures are connected by a kind of entailment links, which unite them. Therefore, we can define the entailment links as a set, which contains all previous determining links, as follows:

D V-2: If **T**(DissF) = ⟨$\mathbf{M_p}$(DissF), **M**(DissF), $\mathbf{M_{pp}}$(DissF), **I**(DissF)⟩ and T(CNT on *dACC*) = ⟨$\mathbf{M_p}$(CNT on *dACC*), **M**(CNT on *dACC*), $\mathbf{M_{pp}}$(CNT on *dACC*), **I**(CNT on *dACC*)⟩ then there exist echelon subsets e1, e2 and the entailment links between them E iff there exist x_1, x_2, λ_1, λ_2, λ_3, λ_4, λ_5 according to D V-8 such that:

(1) x_1 = ⟨Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, attidiff, imp, reward⟩ ∈ $\mathbf{M_p}$(DissF)

(Let x_1 be a potential model of the forced compliance dissonance (**M**$_p$(DissF)))

(2) x_2 = ⟨ N, Act, T, IN, IR, AV, O, I, CW, Cause ⟩ ∈ **M**$_p$(CNT on *dACC*)

(Let x_2 be a potential model of the computational neuroscientific Theory (**M**$_p$(CNT on *dACC*)))

(3) $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$

(Let Λ be a set of determining links between x_1 and x_2)

(4) e_1 = ⟨Time, Cognition, Forcecom, Disscog, Conscog⟩ ∈ an echelon subset of **M**$_p$(DissF) connected to **M**$_p$(CNT on *dACC*) with respect to Λ.

(Therefore, e_1 is an echelon partial substructure of **M**$_p$(DissF) concerning the intertheoretical connections between **T**(DissF) and **T**(CNT on *dACC*) with respect to Λ)

(5) e_2 = ⟨T, AV, CW⟩ ∈ **M**$_p$(CNT on *dACC*) with respect to Λ.

(Therefore, e_2 is an echelon partial substructure of **M**$_p$(CNT on *dACC*) concerning the intertheoretical connections between **T**(DissF) and **T**(CNT on *dACC*) with respect to Λ)

(6) $E = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$

(Therefore, E is a set of the entailment links between the echelon partial substructure e_1(DissF) and the echelon partial substructure e_2(CNT on *dACC*))

Determining the interpreting links for the intertheoretical reduction between DissF and CNT can be done by applying VIII-7 and D VIII-8 in BMS, p. 398–400 and the empirical claim of the interpreting links by applying DVIII-9, and D VIII-10 in BMS, p 402–404. Both are on the non-theoretical level and deal with local empirical claims of the intertheoretical reduction. By defining both, we can know which of the **T**-non-theoretical concepts of one or both theories are relevant for this intertheoretical reduction. This

dissertation will only focus on the local empirical claims of the Festinger theory of forced compliance dissonance as the reduced theory; we could do the same thing for the Hawkins-Kandel computational neuroscientific theory (CNT) by using the same procedure as well. The interpreting links and its local empirical claims on the theory of forced compliance dissonance can be defined as follows:

DV-3 : E*(DissF) = $\{l_1, l_2, l_3, l_4, l_5\}$ is a collection of interpretation links, where T(DissF) is interpreted by T(CNT on *dACC*) and f_1 is the local empirical claims of the interpreting links for the reduction of DissF by CNT on *dACC* on the DissF, iff there exist x_1, x_2 , e_1, e_2, E, y_1 such that:

(1) $x_1 = \langle$Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, attidiff, imp, reward$\rangle \in \mathbf{M_p}$(DissF)
(Let x_1 be a potential model of the forced compliance dissonance ($\mathbf{M_p}$(DissF)))

(2) $x_2 = \langle$ N, Act, T, IN, IR, AV, O, I, CW, Cause $\rangle \in \mathbf{M_p}$(CNT on *dACC*)
(Let x_2 be a potential model of the computational neuroscientific Theory ($\mathbf{M_p}$(CNT on *dACC*)))

(3) $e_1 = \langle$Time, Cognition, Forcecom, Disscog, Conscog$\rangle \in$ an echelon subset of $\mathbf{M_p}$(DissF) connected to $\mathbf{M_p}$(CNT on *dACC*)
(Let e_1 be an echelon subset of $\mathbf{M_p}$(DissF) formed by $\lambda 1$–$\lambda 5$ according to D V-2(4))

(4) $e_2 = \langle$ T, AV, CW $\rangle \in \mathbf{M_p}$(CNT on *dACC*)
(Let e_2 be an echelon subset of $\mathbf{M_p}$(CNT on *dACC*) formed by $\lambda 1$–$\lambda 5$ according to D V-2(5))

(5) $E = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\} \in$ the set of entailment links
(Let E be a set of entailment links between e_1 and e_2)

(6) $y_1 = \langle$Time, Rawcog, Cognition, Disscog, Conscog, pairimp, Forcecom, attidiff, imp, reward$\rangle \in \mathbf{M_{pp}}$(DissF)

(Let y_1 be a partial potential model of the forced compliance dissonance ($\mathbf{M_p}$(DissF)))

(7) $l_1 = \{\langle t_j, time_i\rangle \mid T(t_j) \wedge Time(time_i) \rightarrow R(t_j, time_i)$, where R(x,y) means x interpreting y and R is bijective.

(l_1 is an interpreting link that connects both concepts of Time from $\mathbf{M_{pp}}$(DissF) as identical to T from $\mathbf{M_p}$ (CNT on *dACC*))

(8) $l_2 = \{\langle (cw, aw)_j , cognition_i\rangle \mid (CW, AW)_j \wedge Cognition(cognition_i) \rightarrow R((cw, aw)_j , cognition_i)\}$, where the relation between CW and AV is according to $\forall n_k,n_l \in N: AV(n_k,t) = CW(n_l,n_k).O(n_l)\ I(n_k)\ AV(n_k, t\text{-}1)$.

(l_2 is an interpreting link that interprets the concept of cognition from $\mathbf{M_{pp}}$(DissF) as a function of AV and CW from $\mathbf{M_p}$ (CNT on *dACC*))

(9) $l_3 = \{\langle (cw, aw)_j, forcecom_i \rangle \mid (CW, AW)_j \wedge Forcecom(forcecom_i) \rightarrow R((cw, aw)_j, forcecom_i)\}$, where the relation between CW and AV is according to $\forall n_k,n_l \in N: AV(n_k,t) = CW(n_l,n_k).O(n_l)\ I(n_k)\ AV(n_k, t\text{-}1)$.

(l_3 is an interpreting link that interprets the concept of forcecom from $\mathbf{M_{pp}}$(DissF) as a function of AV and CW from $\mathbf{M_p}$ (CNT on *dACC*))

(10) $l_4 = \{\langle (cw, cw)_j, dc_i\rangle \mid (CW, CW)_j \wedge Disscog(dc_i) \rightarrow R((cw, cw)_j, dc_i)\}$, where (CW,CW) is according to the following function: when for all neurons in the network at a certain period of time $\Sigma_i^m ||CW(t_i) - CW(t_{i+1})||$ is big.

(l_4 is an interpreting link that interprets the concept of Disscog from $\mathbf{M_{pp}}$(DissF) as a function of CW and CW over time from $\mathbf{M_p}$ (CNT on *dACC*))

(11) $l_5 = \{\langle (cw, cw)_j, {}_cc_i\rangle \mid (CW, CW)_j \wedge Conscog(cc_i) \rightarrow R((cw, cw)_j, cc_i)\}$, where (CW,CW) is according to the following function: when for all neurons in the network at a certain period of time $\Sigma_i^m ||CW(t_i) - CW(t_{i+1})||$ is small. (l_5 is an interpreting link that interprets the concept of Conscog from $\mathbf{M_{pp}}$(DissF) as a function of the cross product of CW with itself over time from $\mathbf{M_p}$ (CNT on *dACC*))

(12) f_1 = ⟨Time, Cognition, Disscog, Conscog, Forcecom⟩ ∈ an echelon subset of $\mathbf{M_{pp}}$(DissF) by applying function r*: $e_1 \rightarrow f_1$, that maps E from $\mathbf{M_p}$(DissF) into $\mathbf{M_{pp}}$(DissF).
(f_1 is a set of the empirical claims of the interpreting links for the reduction of DissF by CNT on *dACC*)

The local empirical claims of this intertheoretical reduction f_1 show that on the side of the Festinger theory, all relevant cognitions and their interactions happening in time in DissF, as psychological phenomena, are interpreted as the interaction between the activation and the connections weight in the neurons of a specific part of the brain by the CNT. In the case of van Veen et al.'s research, through the interpreting links the phenomena of cognitive dissonance or cognitive consonance because of forced compliance are understood as the interaction between the neurons' activation and the connection weight of neural networks in the *dorsal Anterior Cingulate Cortex (dACC)* according to the computational neuroscientific theory. The cognitive dissonances within the dissonance group can be associated with high activity of the neural network in the brain to adjust synaptic connections to restore consonance; this is measured by fMRI.

Theoretically, we could also generalize the intertheoretical reduction to the Festinger theory of cognitive dissonance (DissB) by omitting the special rules that create DissF. The only element of the echelon partial substructure of $\mathbf{M_p}$(DissF), which determines the intertheoretical specialization relation between DissF and DissB, is *Forcecom.* Since *Forcecom* is a subset of *Cognition*, the entailment links and the interpreting links can be generalized by omitting *Forcecom*.

A Summary of This Chapter and Some Reflections. This chapter presents a structuralist model to analyze the intertheoretical connections between a specialization of the Festinger theory of cognitive dissonance (called forced compliance dissonance (DissF)) and the computational neuroscientific theory (CNT) in the case of the *dorsal Anterior Cingulate Cortex (dACC)* in Van Veen et. al.'s research program. This kind of intertheoretical relation is an intertheoretical reduction; in this case, only the echelon partial subset of $\mathbf{M_p}$(DissF) is entirely connected to the echelon partial subset of $\mathbf{M_p}$(CNT).

The modeling is done by taking the following steps: (1) We identify the set of determining links (Λ) between both theories, which connect several basic terms of both theories. (2) We characterize the echelon subsets (e) of $\mathbf{M_p}$ of both theories. Based on the second step, we can denote the set of determining links as the set of the entailment links between both echelon subsets (E). (3) Now, we can deal with the non-theoretical level of the intertheoretical connection. In this level, the empirical claims or intended applications of the intertheoretical connections are local on one of both theories according to the concept of theory-holon. Therefore, we have to choose which theory we consider as the local one according to our goal or focus. Supposed, we want to explain the dissonance and consonance according to the DissF in the term of activity of the neurons according to the CNT. Hence, we determined the forced compliance dissonance to be the local theory. We define the interpreting links between both theories, where DissF is interpreted by CNT. We determined the set of interpreting links that connect a part of $\mathbf{M_{pp}}$(DissF) to the echelon subset of $\mathbf{M_p}$(CNT on dACC) by this requirement: $\{l_i = \langle y, x'\rangle |\ x' \in \mathbf{M_p}$(CNT on dACC), $y \in \mathbf{M_{pp}}$(DissF) and there is $x \in \mathbf{M_p}$(DissF) such that $\langle x,x'\rangle \in$(DissF, CNT on dACC) and r*(DissF)=y$\}$ – The function r*(DissF)=y projects the echelon subset of $\mathbf{M_p}$(DissF) to $\mathbf{M_{pp}}$(DissF). And (4) the result of the projection is the class f that is the set of

the empirical claims of our intertheoretical connection. This class f is also the echelon subset of $\mathbf{M_{pp}}$(DissF) as our interpreted theory concerning this intertheoretical reduction. The model can be graphically presented as follows:

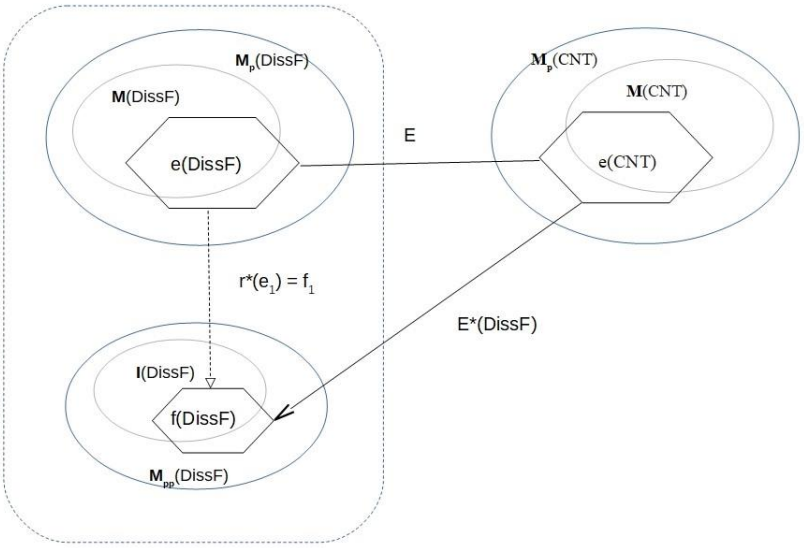


Figure 5.2. A structuralist modeling of the intertheoretical reduction between the theory of forced compliance dissonance (DissF) as the reduced theory and the computational neuroscientific theory (CNT) as the reducing theory; e(DissF) is the echelon subset of $\mathbf{M_p}$(DissF) and e(CNT) is the echelon subset of $\mathbf{M_p}$(CNT). Both are connected by a set of entailment links (E). The interpreting links with local intended applications at DissF (E*(DissF)) connect the set of the local empirical claims of this intertheoretical reduction f(DissF) to the echelon partial subset of $\mathbf{M_p}$(CNT). f(DissF) is the echelon partial subset of $\mathbf{M_{pp}}$(DissF).

Through this modeling and analysis, a piece of confirmation of the notion that the psychological phenomena have connections with how the brain works is delivered by confirmation of intertheoretical reduction between the Festinger theory of cognitive dissonance and the Hawkins-Kandel computational neuroscientific theory put in the context of van Veen et al.'s research. The confirmation can be seen as an empirical confirmation as long as both sets of empirical claims of the interpreting links from both

directions (f(DissF) and f(CNT on dACC) – can be modeled in the similar ways similar to DV-3) are not empty sets. It means that interpretations of each theory by another theory have empirical elements. However, in this case, the confirmation is limited for several reasons: (1) both connected theories are only two of many existing theories in both fields. (2) The intertheoretical connections modeled are only one of many possible connections. (3) The experiments itself is not intended to deliver confirmation of intertheoretical connections in question. (4) The explanation power of these intertheoretical connections is limited to the concepts of the theories connected. We need to connect more theories to get better comprehensiveness and confirmations, but the range of phenomena explained will be more specific (limited).

Although we cannot get a very good degree of confirmation – which is not the goal of this analysis – we still get an example that is representative in many scientific and philosophical practices. The structuralist metatheory of science can be applied for modeling, analyzing, and helping to measure the degree of confirmation of our theory combination or to develop measuring method – by specifying the interconnected parts of the connected theories.

Chapter 6
The Structuralist Models of Intertheoretical Connections between the McCulloch-Pitts Neuron and the Rosenblatt Perceptron and between the Festinger Theory of Cognitive Dissonance and the Hopfield Network in the Consonance Model of Simulation

To understand brain processes during cognitive dissonance, cognitive scientists developed simulations of the process of dissonance reduction by implementing a branch of artificial intelligence inspired by the way how neurons in the brain work, namely artificial neural networks. In Chapters 6 and 7, this dissertation will analyze the intertheoretical connections of some simulation of the Festinger theory of cognitive dissonance. Chapter 6 will discuss the intertheoretical connections of a cognitive dissonance simulation called the consonance model, whereas Chapter 7 the connectionist model. Both simulations apply mathematical models of neurons that are connected in a specific network.

The first model of a neuron, which is still relevant in neuroscience today, was developed by Warren S. McCulloch and Walter H. Pitts in 1943. In the computational neuroscientific theory, the neurons work as a logical switch – there are only two possibilities for the condition of the neurons, i.e., fire (excitatory) or not fire (inhibitory). This theory is based on the McCulloch-Pitts model of a neuron. However, to build a simulation by using an artificial neural network, we must move from the McCulloch-Pitts neuron to the Rosenblatt perceptron. Before starting with the consonance model, it will be interesting to model the intertheoretical connections between the neuron's model commonly used in neuroscience and the neuron's model

commonly used in the artificial intelligence, so that we can know how both models of neurons related each other.

This chapter will consist of two parts. The first part will model the intertheoretical relation of both models of a neuron in order to explore similarities and differences of both models as a preparation for our discussion of the simulation of cognitive dissonance. It is a case of historical reduction, where the McCulloch-Pitts neuron is seen as one of the Rosenblatt perceptron's specializations. The second part will discuss a model of intertheoretical connections between the Festinger theory of cognitive dissonance and the Hopfield network according to the consonance model of simulation.

6.1. The Intertheoretical Relation between the McCulloch-Pitts Neuron and the Rosenblatt Perceptron

Historically speaking, the intertheoretical relation between the McCulloch-Pitts neuron and the Rosenblatt perceptron is a diachronic one: namely, an intertheoretical embedding. “In this kind of scientific change, the models of an older theory get embedded (approximately and perhaps not completely) into the models of a newer, more complex theory in such a way that all (or almost all) intended applications of the older theory, whether successful or unsuccessful, become successful intended applications of the newer one” (Moulines, 2014, p. 1509). The McCulloch-Pitts neuron (1943) is the first generation of a neuron's model, whereas the Rosenblatt perceptron (1958) is the second generation, into which the McCulloch-Pitts neuron gets embedded.

However, his dissertation treats them as in a synchronic relationship because modeling the development of the theory of neuron is not relevant for this project. For this project, knowing the difference and similarity between them is more important because they provide a basis for building a justified and meaningful simulation. Therefore, a model of intertheoretical embedding

will not be built here, but a model of the intertheoretical reduction and specialization of both theories. The goal of this modeling is to illustrate two things formally: The first is similarities and dissimilarities between both theories. The second is the advantage(s) of the Rosenblatt perceptron over the McCulloch-Pitts neuron by changing the possibility of input – to a varying degree – and its various options of the activation function(s) in mathematical models. With a learning rule, a perceptron can also behave as a digital automaton like the McCulloch-Pitts model, besides as a statistical automaton that combines all input data and builds a model representing characteristics of the data.

The similarities and differences between both models can be analyzed through how far the potential models of both theory-elements can be connected by reduction and specialization relations. Here is presented a mathematical model of how the Rosenblatt perceptron reduces to McCulloch-Pitts neuron, which becomes its specialization. Not all elements of potential models of McCulloch-Pitts neuron ($\mathbf{M_p}$(MCP-N)) (see D III-10 in Chapter 3 above) are connected to the elements of potential models of the Rosenblatt Perceptron ($\mathbf{M_p}$(RP)) (see D III-13 in Chapter 3 above). Therefore, modeling the intertheoretical reduction of both theories requires identification of determining links, which connects concepts to concepts in both theories' potential models. The determining links between the theory-element of McCulloch-Pitts neuron **T**(MCP-N) and the theory-element of the Rosenblatt perceptron **T**(RP) can be specified as follows:

DVI-1: If **T**(MCP-N) = $\langle \mathbf{M_p}$(MCP-N), **M**(MCP-N), $\mathbf{M_{pp}}$(MCP-N), I(MCP-N)$\rangle$ and **T**(RP) = $\langle \mathbf{M_p}$(RP), **M**(RP), $\mathbf{M_{pp}}$(RP), **I**(RP)$\rangle$, then there exist λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, λ_7, λ_8, λ_9, λ_{10}, L, e_1, e_2, E between **T**(MCP-N) and **T**(RP) iff there exist x, x' such that:

(1) $x = \langle N, N_0, IR, IN, B, C, W_0, W, Inp, Outp, fnet, fact, fout \rangle \in \mathbf{M_p}$(RP)
(Let x be a potential model of the Rosenblatt Perceptron)

(2) $x' = \langle N', N_0', T, IR, IN, \theta, C', W', Inp', Outp', fnet', fact', fout' \rangle \in \mathbf{M_p}$(MCP-N)

(Let x' be a potential model of the McCulloch-Pitts Neuron)

(3) $\lambda_1 \subset N \times N'$, where a = b, for $a \in N$ and $b \in N'$

(λ_1 is a determining link that connects the set of neurons (*N*) in the potential model of the Rosenblatt perceptron (x) and the set of neurons (*N'*) in the potential model of the McCulloch-Pitts neuron (x'). λ_1 is bijective)

(4) $\lambda_2 \subset N_0 \times N_0'$, where a = b, for $a \in N_0$ and $b \in N_0'$

(λ_2 is a determining link that connects the set of input neurons (N_0) in the potential model of the Rosenblatt perceptron (x) and the set of input neurons (N_0') in the potential model of the McCulloch-Pitts neuron (x'). λ_2 is bijective)

(5) $\lambda_3 \subset C \times C'$, where a = b, for $a \in C$ and $b \in C'$

(λ_3 is a determining link that connects the set of connections between neurons (*C*) in the potential model of the Rosenblatt perceptron (x) and the set of connections between neurons (*C'*) in the potential model of the McCulloch-Pitts neuron (x'). λ_3 is bijective)

(6) $\lambda_4 \subset W \times W'$, where a = b, for $a \in W$ and $b \in W'$

(λ_4 is a determining link that connects the set of connection weights (*W*) in the potential model of the Rosenblatt perceptron (x) and the set of connection weights (*W'*) in the potential model of the McCulloch-Pitts neuron (x'). λ_4 is bijective)

(7) $\lambda_5 \subset (B \times W_0) \times \theta$, where $a \cdot c \supset b$, for $a \in B$, $c \in W_0$, and $b \in \theta$, where $B, W_0 \in IR$ and θ is a constant.

(λ_5 is a determining link that connects the relation between the bias of the neurons (*B*) and its connection weight (W_0) in the potential model of the Rosenblatt perceptron (x) to the set of a threshold of each

neuron (θ) in the potential model of the McCulloch-Pitts neuron (x'). λ_5 is surjective)

(8) $\lambda_6 \subset Inp \times Inp'$, where $a \supset b$, for $a \in Inp$ and $b \in Inp'$, where $Inp \in IR$ and $Inp' \in \{0,1\}$

(λ_6 is a determining link that connects the set of inputs (*Inp*) in the potential model of the Rosenblatt perceptron (x) and the set of inputs (*Inp'*) in the potential model of the McCulloch-Pitts neuron (x'). λ_6 is surjective because $Inp \supset Inp'$)

(9) $\lambda_7 \subset Outp \times Outp'$, where $a \supset b$, for $a \in Outp$ and $b \in Outp'$, where $Outp \in IR$ and $Outp' \in \{0,1\}$

(λ_7 is a determining link that connects the set of outputs (*Outp*) in the potential model of the Rosenblatt perceptron (x) and the set of outputs (*Outp'*) in the potential model of the McCulloch-Pitts neuron (x'). λ_7 is surjective because $Outp \supset Outp'$)

(10) $\lambda_8 \subset fnet \times fnet'$, where $a = b$, for $a \in fnet$ and $b \in fnet'$

(λ_8 is a determining link that connects the network input function of neurons (*fnet*) in the potential model of the Rosenblatt perceptron (x) and the network input function of neurons (*fnet'*) in the potential model of the McCulloch-Pitts neuron (x'). λ_8 is bijective)

(11) $\lambda_9 \subset fact \times fact'$, where $a \supset b$, for $a \in fact$, and $b \in fact'$, where $fact \in IR$ and $fact' \in \{0,1\}$

(λ_9 is a determining link that connects the activation function of neurons (*fact*) in the potential model of the Rosenblatt perceptron (x) and the activation function of neurons (*fact'*) in the potential model of the McCulloch-Pitts neuron (x'). λ_9 is surjective because $fact \supset fact'$)

(12) $\lambda_{10} \subset fout \times fout'$, where $a \supset b$, for $a \in fout$, and $b \in fout'$, where $fout \in IR$ and $fout' \in \{0,1\}$

(λ_{10} is a determining link that connects the output function of neurons (*fout*) in the potential model of the Rosenblatt perceptron (x) and the

output function of neurons (*fout'*) in the potential model of the McCulloch-Pitts neuron (x'). λ_{10} is surjective because fout $\supset$ fout')

(13) $\Lambda = \{ \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8, \lambda_9, \lambda_{10}\}$

(Λ is a collection of determining links that connect the concepts of both theories)

(14) $e_1 = \langle N, N_0, IR, IN, B, C, W_0, W, Inp, Outp, fnet, fact, fout \rangle \in$ an echelon partial subset of $\mathbf{M_p}$(RP)

(e_1 is an echelon partial subset of the potential model of the Rosenblatt perceptron)

(15) $e_2 = \langle N', N_0', IR, IN, \theta, C', W', Inp', Outp', fnet', fact', fout' \rangle \in$ an echelon partial subset of $\mathbf{M_p}$(MCP-N)

(e_2 is an echelon partial subset of the potential model of the McCulloch-Pitts neuron)

(16) $E = \{ \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8, \lambda_9, \lambda_{10}\}$

(E is a set of entailment links that connect the echelon partial subset of the potential model of the Rosenblatt perceptron and the echelon partial subset of the potential model of the McCulloch-Pitts neuron)

In this model, theoretical similarities and dissimilarities between the McCulloch-Pitts neuron and the Rosenblatt perceptron can be identified by similarities and differences of terms or concepts of both theories connected by determining links connecting sets of the same terms. The theoretical similarities between the McCulloch-Pitts neuron and the Rosenblatt perceptron are determined through the following concepts:

(1) The concepts of neurons (*N*), neuron input (N_0), connections (*C*), and connection weight(*W*) and network input function (*fnet*) of both models are similar.

(2) The concept of the threshold of the McCulloch-Pitts neuron is a true subset of a multiplication of the bias (*B*) and its connection weight (W_0) in the Rosenblatt perceptron.
(3) The terms input (*Inp*), output (*Outp*), and output function (*fout*) are equivalent *as long as* these terms of the Rosenblatt perceptron cover only 0 and 1.
(4) The concept of the activation function (*fact*) of both models is equivalent *as long as* they implement the same activation function, namely the linear function.

The theoretical dissimilarities can be determined by the intertheoretical reduction of both models that can be identified by the two following conditions. The first condition is that there is a term with no determining link at all, namely the concept of time (*T*) in the McCulloch-Pitts neuron. It "reduces" applicability of the Rosenblatt perceptron. The second condition is that there exist some determining links that are not bijective but surjective. These links show not only how far their similarities but also reduce the applicability of the Rosenblatt perceptron. The concepts of *Inp, Outp, fact,* and *fout* of both models are not entirely similar, because these concepts of McCulloch-Pitts neurons are only true subsets of the respective concepts of the Rosenblatt perceptron.

Modeling the empirical reduction between both models of the artificial neuron can be done by projecting the determining intertheoretical connections between both theories onto their **T**-non-theoretical level. It can be done by omitting the determining links that connect the **T**-theoretical elements of both potential models. As we see in Chapter 2, the **T**-theoretical elements in the potential models of the McCulloch-Pitts neuron are their *fnet, fact,* and *fout*, whereas the **T**-theoretical elements in the potential models of the Rosenblatt perceptron are also their *fnet, fact,* and *fout*. Here, we speak about theory-nets and not about theory-holon because both theories are in relatives. Therefore, defining local empirical claims of the intertheoretical

relation is not required. The determining links that represent the empirical similarities – and dissimilarities – of both theories are as follows:

DVI-2: Λ^* is a collection of determining links that represents the empirical equivalence between the theory-element of McCulloch-Pitts neuron and the theory-element of the Rosenblatt perceptron and f_1 is an echelon partial subset of the partial potential model of the McCulloch-Pitts neuron iff there exists x, x', Λ, y, y' such that:

(1) $x = \langle N, N_0, IR, IN, B, C, W_0, W, Inp, Outp, fnet, fact, fout \rangle \in \mathbf{M_p}(RP)$
(Let x be a potential model of the Rosenblatt perceptron)

(2) $x' = \langle N', N_0', T, IR, IN, \theta, C', W', Inp', Outp', fnet', fact', fout' \rangle \in \mathbf{M_p}(MCP\text{-}N)$
(Let x' be a potential model of the McCulloch-Pitts neuron)

(3) $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8, \lambda_9, \lambda_{10}\}$
(Let Λ be a collection of the determining links between both potential models)

(4) fnet, fact, fout are **T**-theoretic
(The **T**-theoretical elements in the potential model of the Rosenblatt perceptron)

(5) fnet', fact', fout' are **T**-theoretic
(The **T**-theoretical elements in the potential model of the McCulloch-Pitts neuron)

(6) $y = \langle N, N_0, IR, IN, B, C, W_0, W, Inp, Outp \rangle \in \mathbf{M_{pp}}(RP)$
(y is a partial potential model of the Rosenblatt perceptron)

(7) $y' = \langle N', N_0', T, IR, IN, \theta, C', W', Inp', Outp' \rangle \in \mathbf{M_{pp}}(MCP\text{-}N)$
(y' is a partial potential model of the McCulloch-Pitts neuron)

(8) $\Lambda^* = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7\}$
(Λ^* is be a collection of the determining links of both partial potential models)

(9) $f_1 = \langle N, N_0, IR, IN, B, C, W_0, W, Inp, Outp\rangle \in$ echelon partial subset of $\mathbf{M_{pp}}$(RP)

(f_1 is an echelon partial subset of the partial potential model of the Rosenblatt perceptron)

(10) $f_2 = \langle N', N_0', IR, IN, \theta, C', W', Inp, Outp\rangle \in$ echelon partial subset of $\mathbf{M_{pp}}$(MCP-N)

(f_2 is an echelon partial subset of the partial potential model of the McCulloch-Pitts Neuron)

Through this model, the intertheoretical reduction and specialization can be illustrated through the following figure:

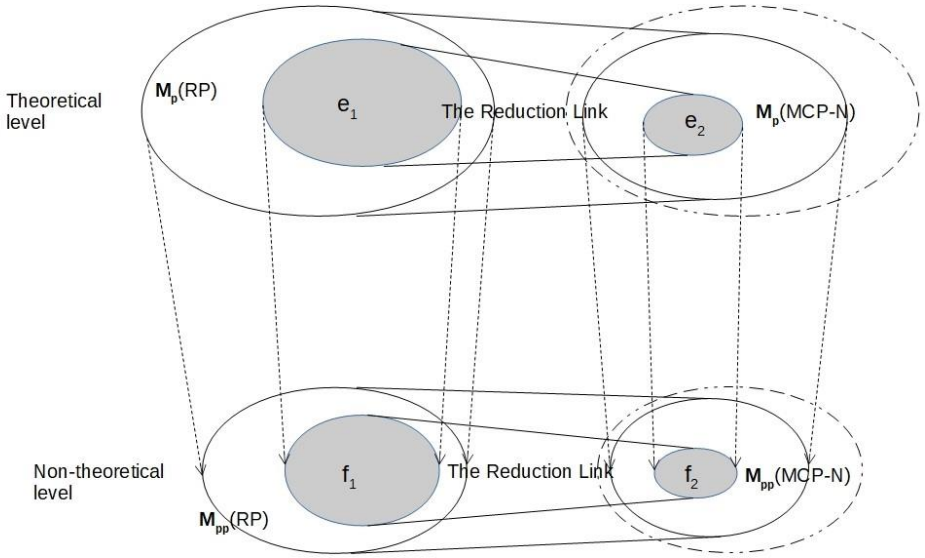


Figure 6.1. The reduction link reduces the theory-element of the Rosenblatt perceptron (**T**(RP)) to the theory-element of the McCulloch-Pitts neuron (**T**(MCP-N)) and as its result **T**(MCP-N) becomes a specialization of **T**(RP)

The relation of both theories also matches the requirements of specialization relation in DIV-1.

(1) $\mathbf{M_p}$(MCP-N) $\subset$ $\mathbf{M_p}$(RP), $\mathbf{M_{pp}}$(MCP-N) $\subset$ $\mathbf{M_{pp}}$(RP)

(2) **M**(MCP-N) ⊂ **M**(RP), **GL**(MCP-N) ⊂ **GL**(RP), and **I**(MCP-N) ⊂ **I**(RP).

In the next part of this chapter and the next chapter, the structuralist metatheory of science is applied not for modeling empirical theories, but for modeling artificial neural networks for simulation. Some fellow structuralists may not agree with this research because the structuralist theory is particularly suitable for modeling empirical theories, whereas the artificial neural network is not an empirical model but a mathematical model. This dissertation sees that the application STS is not only limited to modeling empirical theories but also for modeling non-empirical scientific theories, such as artificial neural networks. This application is based on some considerations as follows: First, $\mathbf{M_{pp}}$ contains not only empirical or observational terms but also non-observational terms, i.e., terms from other theories. (2) Es is possible that terms x_1,.., x_{10} in theory **T** are not empirical, but they are related or taken over analogically from the terms x_1',… x_{10}' from **T'**, which are empirical. It is precisely the case of Rosenblatt perceptron and the McCulloch-Pitts neuron. (3) For building the simulation, the **T**-theoretical terms of the theory element of the McCulloch-Pitts neuron **T**(MCP-N) such as *fnet*, *fact*, and *fout* can be taken over as the **T**-theoretical terms of the theory-element of the Rosenblatt perceptron **T**(RP) because both theory-elements are in the same theory-net (see DVIII-3, BMS, p. 392).

6.2. The Consonance Model

The consonance model is developed by Thomas R. Shultz and Mark R. Lepper. This model simulates the mind as a mechanism that maintains some equilibrium and the dissonance reduction as a solving of the constraint satisfaction problem among someone's beliefs and behaviors. "The model is based on the idea that dissonance reduction can be viewed as a constraint

satisfaction problem. … the motive to seek cognitive consistency postulated by dissonance theory and related models … can be seen as imposing constraints on the beliefs and attitudes that an individual holds simultaneously …. Such problems can be solved by the simultaneous satisfaction of many soft constraints that can vary in their relative importance. Soft, as opposed to hard, constraints are those that are desirable, but not essential, to satisfy" (Shultz and Lepper, 1996, p. 220). In this model, the network is being used to simulate "the subject's representation of the situation created, or psychological problem posed, by the experimental settings in the classic cognitive dissonance paradigms" (Shultz and Lepper, 1996, p. 220). Shultz and Lepper implemented the Hopfield network because it can be used to solve complex optimization problems, where network states with low energy levels represent optimal solutions.

The basic idea of this simulation is as follows: The units (or neurons) represent the cognition involved in arousal of dissonance and reduction of dissonance. The activation of units represents the direction and strength of an individual's beliefs and attitudes. The units can differ in their resistance to change according to differences related to the fact that cognition may be supported by other cognition or anchored in reality. Connection weights represent psychologically causal implications among the individual's beliefs and attitudes. Therefore, they can be either excitatory (+), inhibitory (-), or psychologically irrelevant (0). We can initially adapt unit activations and connection weights, depending on the paradigm, according to the different conditions of a single experiment (Shultz and Lepper, 1996, p. 220).

In this simulation, increasing consonance corresponds to the process of reducing dissonance or striving for consistency among individual's beliefs and attitudes. Shultz and Lepper defined consonance as "the degree to which similarly active units are linked by excitatory (+) weights and differently active units or inactive units are linked by inhibitory (-) weights" (Shultz and Lepper, 2009, p. 238). In the consonance model of simulation, Shultz and

Lepper implement the Hopfield network with some modifications. “Hopfield worked out the mathematics for solving constraint satisfaction problems in parallel networks. Maximizing the consonance (or goodness) of any pair of connected units depends on the sign of the connection between them …. If connected by a positive weight, both units of the pair should be active to maximize consonance. With a negative weight, consonance is maximized when the two units are not both active; that is, when both are inactive, or only one is active. Activation will change over time cycles so as to satisfy the various constraints and increase consonance” (Shultz and Lepper, 1996, p. 220).

This consonance model works according to the following computational rule (Shultz and Lepper, 1996, pp. 220–221, 2009, pp. 239–241): The consonance contributed by a particular unit i can be defined as follows:

$$\text{consonance}_i = \Sigma_j \, w_{ij} \, a_i a_j \qquad (1)$$

Where w_{ij} is the connection weight between units i and j, a_i is the activation of receiving unit i, and a_j is the activation of the sending unit j. The network's consonance is defined as a sum of consonance of all receiving units in the network:

$$\text{consonance}_n = \Sigma_i \Sigma_j \, w_{ij} \, a_i a_j \qquad (2)$$

The Hopfield network is a recurrent neural network. Therefore, all units are both a receiving unit and a sending unit. Activation of all units is propagated over time in the network according to the following rules for updating the unit's activation:

$$a_i(t+1) = a_i(t) + \text{net}_i[\text{ceiling} - a_i(t)], \text{ when } \text{net}_i \geq 0, \qquad (3)$$

$$a_i(t+1) = a_i(t) + \text{net}_i[a_i(t) - \text{floor}], \text{ when } \text{net}_i < 0, \qquad (4)$$

where $a_i(t+1)$ is the activation of receiving unit i at time t+1, $a_i(t)$ is the activation of unit i at time t, ceiling is the maximal level of unit activation,

floor is the minimal level of unit activation, and net_i is the scaled net input to unit i, which can be defined as:

$$net_i = resist_i \Sigma_j w_{ij} a_j \quad (5)$$

The parameter $resist_i$ indicates the resistance of receiving unit i to having its activation changed. Smaller values of this parameter indicate greater resistance because smaller values mean less impact of the network input. The network input of a unit is updated by the sum of the products of connection weight and activation of their sending unit over the time of the simulation.

The implementation of the consonance model for cognitive dissonance simulation requires six theoretical principles for mapping the consonance model onto the dissonance theory (Shultz and Lepper, 2009, p. 245). These principles constrain the design of networks representing the (general) conditions of each experiment. "To varying degrees, these theoretical principles were specified in classical dissonance theory. Additional specifications, where necessary, are supplied by the consonance model. Each theoretical principle governs the design of all simulations with the consonance model" (Shultz and Lepper, 1996, p. 221). These principles are as follows:

Principle 1: Representation of cognition. Although cognitions are the basic elements of the theory of cognitive dissonance, they are not fully specified in this theory. "They can be both beliefs and evaluations (i.e., attitudes), such cognitions could be assumed to vary in both direction and strength. The positive direction could represent that something is either believed to be true or is favorably evaluated. …, the negative direction could represent that something is either believed to be false or is negatively evaluated. Strength is the degree to which something is believed to be true or false or evaluated positively or negatively" (Shultz and Lepper, 1996, p. 221). Therefore, in this network, each cognition is represented by the net activation of a pair of negatively connected units – one unit represents a positive

direction, and the other represents the negative direction. Respectively, a net activation for the cognition is the difference between the activation of the positive unit and the activation of the negative unit. The neurological and computational plausibility for this has been reviewed by Anderson (1995, pp. 150–152), where the neurons can be found in two organized groups, which are inhibitory and excitatory at the same time related to some specific input. The activation range for the positive neurons is also higher than the one for the negative neuron. For mimicking this fact, Shultz and Lepper set ceiling activation parameter to 1 for units representing positive aspects of cognition, and to 0.5 for units representing the negative aspect of cognition. The floor parameter is set to 0 for both types. For undertaking the simulation, the initial activation is generally set by default to the value 0.5 for high – strongly believed – and 0.1 for low – weakly believed.

This representation allows some degree of ambivalence in cognition, such as something both liked and disliked. The inhibitory connections between the two poles tend to discourage such ambivalence. However, relatively persistent ambivalence can be produced if both the positive and negative units for a cognition receive strong support from other cognitions. Such ambivalence creates dissonance, as explained in the next principle.

Principle 2: Relationships among cognitions. In the consonance model, cognitions are connected to other cognitions based on their causal implications to form a network, which represents a person's relevant beliefs and attitudes regarding a particular experimental situation. A negative implication is represented by an inhibitory (-) weight between two cognitions; a positive implication is represented by an excitatory (+) weight. Connection weights range from -1 to 1, with 0 representing a lack of causal relation. If two cognitions are positively related, their positive poles are connected to an excitatory weight; it is similar for their negative pole. Inhibitory (-) weights connect the positive pole of one cognition with the negative pole of another cognition and vice versa (Figure 4.3.A). If two cognitions are negatively

related, their positive poles are connected to the inhibitory weights; it will be similar for their negative poles. The excitatory weights connect the positive pole of one cognition with the negative pole of another cognition and vice versa (Figure 4.3.B). For both cases, each unit has an inhibitory self-connection, and all connection weights are bidirectional. In this simulation, the initial connection weight is set by default to 0.5 for a strong connection, and for a weak connection to 0.1 by default.

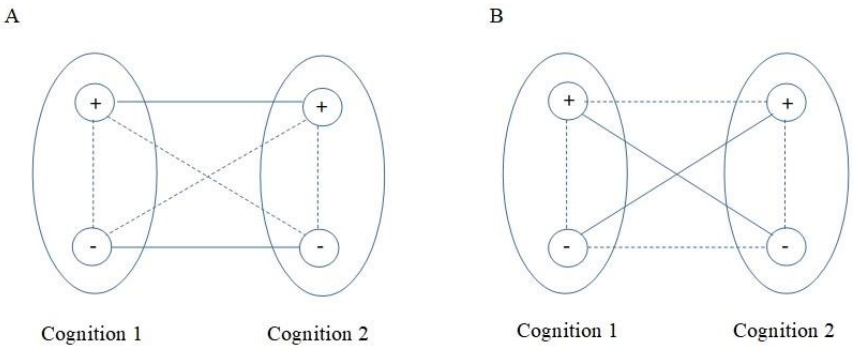


Figure 6.2. Any two cognitions can be connected positively (as shown in Figure A), negatively (as shown in Figure B), or can be unrelated. In this figure, positive connection weights are symbolized by the solid lines and negative connection weights by the dashed lines. Each connection is symbolized by an ellipse drawn around the positive and negative poles of the cognition. (Source: Shultz and Lepper, 1996, p. 222)

Principle 3: Magnitude of dissonance. According to cognitive dissonance theory, the total amount of dissonance is a function of the ratio of dissonance cognitions to all relevant cognitions (dissonant plus consonant cognition) with cognitions and relations weighted for their importance to a person. In the consonance model, the total consonance of the network is represented by Equation 2 above, which is defined by the triple products of sending activation, receiving activation, and connection weight summed.

Total dissonance is formally defined as the negative of total consonance divided by r, the number of non-zero inter-cognition relations in the network.

$$\text{Dissonance} = - \Sigma_i \Sigma_j w_{ij} a_i a_j / r \quad (6)$$

Dividing by r has the goal of standardizing dissonance across networks of different size and connection density by controlling the number of relevant relations. The self-connection of units is excluded from this computation.

The following things are worth mentioning: (1) The definition of dissonance is analogous to the Hopfield network's definition of energy. (2) The total consonance is a triple product of sending activation, receiving activation, and connecting weight summed. In term of believability, the larger their numeric value, the larger is their impact on consonance. (3) Irrelevant cognition that is connected by a weight of 0 contributes nothing to consonance. (4) This definition of dissonance offers some advantages over the definition given by Festinger because it is mathematically formalized. Therefore, it provides: (a) an easy application to complex belief structures, (b) the possibility of measurement of the amount of dissonance in each inter-cognition relation and of ambivalence within the cognition, and (c) variations according to all possible individual relations, whether consonance, dissonance, or mixed.

Principle 4: Dissonance reduction. According to the theory of cognitive dissonance, people have a strong tendency to reduce cognitive dissonance, although this is not always successful. In the consonance model, networks tend to settle into more stable, lower dissonance states by updating all unit activations according to equations 3–5 above. Shultz and Lepper set two parameters that affect the process of dissonance reduction, namely *cap* and *rand%*: "A *cap* parameter, with a default of -0.5, corresponds to the connection between each unit and itself, w_{ii}, and prevents activations from reaching the activation ceiling. … The *rand%* parameter provides a means of globally testing the robustness of the results obtained in simulations in the

face of variations in the specific numerical values used to instantiate key variables. At the start of each network run, connection weights, resistances, and initial activations all are randomized by adding or subtracting a random proportion of their initial amounts. The *rand%* parameter specifies the proportion range in which these additional and subtraction are selected under a random uniform distribution. Ordinarily, we use small (0.1), medium (0.5), and large (1.0) levels of *rand%*. The randomizing network values in this way increases psychological realism because not every person can be expected to share exactly the same parameter values" (Shultz and Lepper, 2009, p. 244). This *rand%* parameter randomizes these basic variables as follows:

$$y = x \pm \{\text{random}(\text{absolute}[x \times \text{rand}\%])\} \quad (7)$$

The randomization of weight values violates the symmetry of connection weights as assumed by Hopfield (1982, 1984), so that $w_{ij} \neq w_{ji}$, and thus makes network solutions less stable, meaning that outcomes are more variable. Related to this randomization in the Hopfield network, Shultz and Lepper write, "that violations of the symmetry assumption increased memory errors and instability in network solutions to memory retrieval problems. Such results may also correspond to natural psychological variation" (Shultz and Lepper, 1996, p. 224).

Principle 5: Changes in Cognitions. According to Festinger (1957), dissonance can be reduced by decreasing the number or importance of dissonant relations, by increasing the number or importance of consonant relations, or by a combination of these factors. Changing the importance of dissonance can be done by changing evaluations, beliefs, and implications among them. In the theory of cognitive dissonance, there are different degrees of resistance to change among the cognitions; the cognitions most likely to change are the least resistant to change. For example, beliefs are more resistant to change than evaluations or attitudes. "Resistance stems from the possible creations of new dissonance because of relations with other

cognitions; from cognition that is anchored in reality, and from the difficulty of changing aspects of reality" (Shultz and Lepper, 1996, p. 224). According to dissonance experiments, dissonance is reduced by changes in evaluations, not by changes in belief about the salient event presented nor by changes in implications among cognitions (Shultz and Lepper, 1996, p. 224). For simulating this fact, these connection weights in the consonance model are not allowed to change, but cognition unit activations change over time as dissonance is reduced. The resistance parameter ($Resist_i$) simulates this phenomenon in equation five above. By default, this parameter is set to the values of 0.5 for low resistance and of 0.1 for high resistance.

Principle 6: Importance of Dissonance. In their model, Shultz and Lepper implement the idea of the importance of dissonance by multiplying all connection weights and unit activations with a certain number at the beginning of every simulation. "An importance parameter multiplies all connection weights and unit activation at the start of each run, before the initial randomizations described under Principle 4. Typically, we use values of 1.0 in control conditions, 0.5 for conditions that lessen the importance of a dissonance situation, and 1.5 for conditions that enhance the importance of a dissonance situation. … The precise values used are somewhat arbitrary, but it is important, that they differ substantially to generate different results" (Shultz and Lepper, 2009, p. 245).

The simulation, according to this consonance model, is done by using *a generic consonance model*. In this model, cognitions are categorized into three categories: behaviors, justifications, and evaluations. In several simulations that implement this model, "the consonance model enables specification of each of the relevant cognitions, including their type and their initial activations, and of the relations among cognitions. Different dissonance experiments require different instantiations of this generic network because they involve different particular types of cognition, with differing particular initial activation values, and particular implications

among cognitions. … evaluation cognition is given low resistance, whereas other cognitions types (about behavior and justifications) are given high resistance" (Shultz and Lepper, 1996, p. 224).

6.3. The Structuralist Model of the Intertheoretical Connections between Festinger's Theory of Cognitive Dissonance and the Hopfield Network for the Consonance Model

The consonance model is a simulation of dissonance reduction through an artificial recurrent neural network, namely the Hopfield network. The motivation of this simulation is to simulate how neural activities in the brain can produce such psychological processes, namely reducing the dissonance at hand.

Building a model of intertheoretical connections of the consonant model requires some modifications of the Hopfield network. The first modification is related to the first principle in the consonant model, namely representation of cognition: "They can be both beliefs and evaluations (i.e., attitudes), such cognition could be assumed to vary in both direction and strength. The positive direction could represent that something is either believed to be true or is favorably evaluated. …, the negative direction could represent that something is either believed to be false or is negatively evaluated. Strength is the degree to which something is believed to be true or false or evaluated positively or negatively" (Shultz and Lepper, 1996, p. 221). It requires to modify the Hopfield network by adding a differentiation between neurons and the relation between those neurons. The first modification is eliminating differentiation among *input neurons, hidden neurons, output neurons, pred(n),* and *succ(n)*, and creating differentiation between neurons (+) and neurons (-) to represent positive and negative valuations.

(1) N^+ is a finite non-empty set of neurons (+).

(2) N^- is a finite non-empty set of neurons (-).

(3) $N^+ \subseteq N$

(4) $N^- \subseteq N$

The second modification is a differentiation among the connections between neurons, which differentiates between pairs of neurons (*Pairs*) representing a cognition, and relations between members of *Pairs* (*Conpairs*) that represent the relation between cognitions. In the consonance model, the cognition is represented by a pair of neurons; the set of *Pairs* represent a set of cognitions. Therefore, the connection between cognitions will be represented by connections between elements of *Pairs*. The set of Conpairs will represent the*se connections between cognitions*.

(1) Pairs $\subseteq N^+$' N^-, all members of Pairs are bijective relations.

(2) Pairs $\subset$ C

(3) Conpairs $\subseteq$ Pairs $\times$ Pairs

(4) C = Pairs $\cup$ Conpairs

For the simulation, ext_n and out_n, are not required anymore; all required is the activation of the pairs of neurons, that represent the valuations of the cognition. Therefore, they are put aside. Because the simulation does not have input and output, the network input net_n will be modified. Defining net_n needs another new parameter that behaves like a learning parameter. Shultz and Lepper call it as $resist_i$ "that indicates the resistance of receiving unit i to having its activation changed. Smaller values of this parameter indicate greater resistance because smaller values mean less impact of the net input" (Shultz and Lepper, 1999, p. 240). The law statement to determine net_i will be $net_i = resist_i \Sigma_j w_{ij} a_j$.

We also need a variable for time *t* to capture the update of the activation of receiving unit over time. The updating rules are as follows:

$a_i(t+1) = a_i(t) + net_i[ceiling - a_i(t)]$, when $net_i \geq 0$,

$a_i(t+1) = a_i(t) + net_i[a_i(t) - floor]$, when $net_i < 0$,

Therefore, we also need two additional terms, *ceiling* and *floor*. Shultz and Lepper set *ceiling* activation parameter to 1 for units representing positive aspects of cognitions, and to 0.5 for units representing negative aspect of cognitions. The *floor* parameter is set to 0 for both types.

The next modification is related to the definition of consonance and dissonance. The term consonance replaces the term of state from the old $\mathbf{M_p}$(HN) (see D III-20 of Chapter 3 above), whereas the term dissonance replaces the term energy E from the old $\mathbf{M_p}$(HN). The consonance contributed by a particular unit i can be defined as follows:

$$\text{consonance}_i = \Sigma_j w_{ij} a_i a_j$$

where w_{ij} is the connection weight between units i and j, a_i is the activation of receiving unit i, and a_j is the activation of the sending unit j. The consonance of the network is defined as a sum of the consonances of all receiving units in the network:

$$\text{consonance}_n = \Sigma_i \Sigma_j w_{ij} a_i a_j$$

The total dissonance is formally defined as the negative of total consonance divided by r, the number of non-zero inter-cognition relations in the network.

$$\text{Dissonance} = - \Sigma_i \Sigma_j w_{ij} a_i a_j / r$$

The last modification is the addition of two parameters that affect the process of dissonance reduction, i.e. *cap* and *rand%*: "A *cap* parameter, with a default of -0.5, corresponds to the connection between each unit and itself, w_{ii}, and prevents activations from reaching the activation ceiling.… The *rand%* parameter provides a means of globally testing the robustness of the results obtained in simulations in the face of variations in the specific numerical values used to instantiate key variables. At the start of each network run, connection weights, resistances, and initial activations are all randomized by adding or subtracting a random proportion of their initial amounts. The *rand%* parameter specifies the proportion range in which these additional and subtraction are selected under a random uniform distribution. Ordinarily, we

use small (0.1), medium (0.5), and large (1.0) levels of *rand%*. Randomizing network values in this way increases psychological realism, because not every person can be expected to share exactly the same parameter values" (Shultz and Lepper, 2009, p. 244). Therefore, the potential models and the actual models of the Hopfield network (D III-20 & DIII-21) in Chapter 3 are modified into DVI-2 and DVI-3 as follows:

DVI-3: x is a potential model of the architecture of the Hopfield network for the consonance model (x ∈ $\mathbf{M_p}$(HN for Consonance)) iff there exist N, T, N^+, N^-, C, Pairs, Conpairs, W, Resist, A, net_n, State, E so that:

(1) X = ⟨N, T, IR, N^+, N^-, C, Pairs, Conpairs, W, Resist, A, Imp, cap, rand%, net_n, ceiling, floor, consonance, Dissonance⟩ ∈ $\mathbf{M_p}$(HN for Consonance) (let x be a potential model of the adapted Hopfield network for the consonance model)

(2) $N \neq \varnothing$ (Let *N* is a non-empty set of neurons)

(3) $T \neq \varnothing$

(*T* is a finite non-empty set of discrete points of time)

(4) $N^+ \subseteq N$

(N^+ as a subset of *N* is a finite non-empty set of neurons, that represent positive valuations of cognition)

(5) $N^- \subseteq N$

(N^- as a subset of *N* is a finite non-empty set of neurons, that represent negative valuations of cognitions)

(6) $C \subseteq N \times N$

(*C* is a finite non-empty set of directed connections between neurons)

(7) Pairs $\subseteq N^+ \times N^-$

(*Pairs* is a relation between two neurons with two opposite valuations that represent a certain cognition. *Pairs* is bijective)

(8) Conpairs $\subseteq$ Pairs×Pairs

(*Conpairs* is a set of connections between *Pairs*, which represents the relationships between cognitions)

(9) $W := C \rightarrow \{-1, 0, 1\}$

(*W* is a function that maps every connection into an element of the set {-1, 0, 1}. This number represents the weight of connections, which represent the characteristics of the relation between the neurons in pairs)

(10) $Resist_i := N \rightarrow IR$

(*Resist* is a function that maps each neuron to a real number that represents the resistance of the neuron i to having its activation changed)

(11) $A := N \rightarrow IR$

(*A* is a function that maps each neuron to a real number that represents the activation of the unit/neuron)

(12) $Imp := IR \rightarrow IR$

(*Imp* is a function that represents the importance parameter. It is implemented on *A* and *W*)

(13) $cap := C \rightarrow IR$

(*cap* is a function that maps each neuron to a real number, which represents the synaptic weight of units with itself. By default, this synaptic weight is set to -0.5)

(14) $rand\% := IR \rightarrow IR$

(*Rand%* is a function of random parameters)

(15) $net_n \subseteq resist \times actn \times W \rightarrow IR$

(*net_n* is a function that calculates the network input of each neuron)

(16) $ceiling := N \rightarrow IR$

(*ceiling* is a function that maps each neuron to a real number as a parameter representing the maximal level of unit activation: This is

set to 1 for units representing positive aspects of cognition, and to 0.5 for units representing the negative aspect of cognition)

(17) floor := $N \rightarrow IR$

(*floor* is a function that maps each neuron to a real number as a parameter representing the minimal level of unit activation set to 0 for both types)

(18) Consonance $\subseteq act_n \times W \times act_n$

(*Consonance* is a relation that represents the state of the Hopfield net)

(19) Dissonance := State $\rightarrow$ IR

(*Dissonance* is a function that represents the magnitude of dissonance)

DVI-3: x is an actual model of the architecture of the Hopfield network for the consonance model ($x \in$ **M**(HN for Consonance)) iff there exist N, C, T, N^+, N^-, Pairs, Conpairs, W, Resist, A, cap, rand%, net_n, ceiling, floor, consonance, Dissonance such that:

(1) $x = \langle N, T, IR, N^+, N^-, C, Pairs, Conpairs, W, Resist, A, Imp, cap, rand\%, net_n, ceiling, floor, consonance, Dissonance\rangle \in$ $\mathbf{M_p}$(HN for Consonance)

(let x be a potential model of the Hopfield network for the consonance model.)

(2) $N = N^+ \cup N^-$

(The set of neurons in the network consists only of neurons with positive valuation and neurons with negative valuation)

(3) $N^+ \cap N^- = \varnothing$

(The intersection of the sets of neurons with positive valuation and the set with negative valuation is an empty set. Therefore, there exist no neuron with both and no valuation(s))

(4) $C = N \times N - \{(n_1, n_2) | \ n_1 \in N \wedge n_2 \in N \rightarrow n_1 = n_2\}$

(*C* is a set of the connections between a neuron and other neurons)

(5) $n_i, n_j \in N, n_i \neq n_j$: $w_{ij} = [-1, 1]$, where $w_{ij} = 0$ means a lack of causal relation.

(The synaptic weight ($w \in W$) has a value between -1 and 1, where $w_{ij} = 0$ means a lack of causal relation.)

(6) For all $n_i, n_j \in N^+$: $w_{ij} =]0, 1]$ and for all $n_i, n_j \in N^-$: $w_{ij} =]0,1]$.

(The synaptic weight ($w \in W$) between two neurons with the same valuation is between 0 and 1. Zero is not included)

(7) For all $n_i \in N^+$, $n_j \in N^-$: $w_{ij} = [-1, 0[$.

(The synaptic weight ($w \in W$) between two neurons with the different valuation is between -1 and 0. Zero is not included)

(8) For all $n_i, n_j \in N$, $n_i = n_j$: $cap_{ij} = cap_{ji} = -0.5$

(The connection weight for intra-neural connections is set -0.5)

(9) For all $x \in A, W$: $x^* = Imp(x)$.

for control condition: $x^* = 1.0x$

for lessened importance: $x^* = 0.5x$

for enhanced importance: $x^* = 1.5x$

("The importance parameter *Imp* multiplies all connection weights and the unit activations at the start of each run, before the initial randomization under [D VI-3 (10)]" (Shultz and Lepper, 1999, p. 245))

(10) For initial $x \in A, W$, Resist: $x_{initial} = x \pm \{random(absolute[x \cdot rand\%])\}$

(*A, W, Resist* are initialized in the beginning of simulation by using a random number according to $y = x \pm \{random(absolute[x \times rand\%])\}$)

(11) For $n_i, n_j \in N$, $resist_i \in Resist$, $w_{ij} \in W$, $i,j = 1, \ldots, k$: $net_i = resist_i \Sigma_1^k w_{ij} a_j$.

(Impact of parameter *Resist* on the net. Smaller values of *Resist* indicate greater resistance because smaller values mean less impact of the net input)

(12) For all $n_i \in N$, $t_j \in T$, j = 0, 1, …, k:

(a) for $t = t_0$:

$a_i(t_{j+1}) = a_{\text{initial } i}(t_j) + net_i[\text{ceiling} - a_{\text{iinitial } i}(t_j)]$, when $net_i \geq 0$,

$a_i(t_{j+1}) = a_{\text{initial } i}(t_j) + net_i[a_{\text{iinitial } i}(t_j) - \text{floor}]$, when $net_i < 0$,

(b) for $t = t_j$, j = 1, …, k:

$a_i(t_{j+1}) = a_i(t_j) + net_i[\text{ceiling} - a_i(t_j)]$, when $net_i \geq 0$,

$a_i(t_{j+1}) = a_i(t_j) + net_i[a_i(t_j) - \text{floor}]$, when $net_i < 0$,

(The activation function during the time considered.)

(13) for all $n_i, n_j \in N$:

(a) $\text{consonance}_i = \Sigma_j w_{ij} a_i a_j$

(b) $\text{consonance}_n = \Sigma_i \Sigma_j w_{ij} a_i a_j$

(The consonance function of the network)

(14) $\text{Dissonance} = - \Sigma_i \Sigma_j w_{ij} a_i a_j / r$

(The dissonance function of the network)

The partial potential models of the Hopfield Network for the consonance model $\mathbf{M_{pp}}$(HN for Consonance) can be modeled by omitting the **T**-theoretical elements of the potential models of the Hopfield network, i.e., *Resist, A, imp cap, rand%, net_n, ceiling, floor, consonance, Dissonance*. These concepts are **T**-theoretical because they are defined without any empirical observation or any other theories. They are determined so that the system behaves naturally. The partial potential models of the Hopfield network for the consonant model can be defined as follows:

DVI-4: y is a partial potential model of the Hopfield Network for the consonance model (y = $\mathbf{M_{pp}}$(HN for Consonance)) iff there exist x such that:

(1) $x = \langle N, C, T, N^+, N^-, \text{Pairs}, \text{Conpairs}, W, \text{Resist}, A, \text{imp}, \text{cap}, \text{rand\%}, net_n, \text{ceiling}, \text{floor}, \text{consonance}, \text{Dissonance} \rangle \in \mathbf{M_p}$(HN for Consonance)

(2) Resist, A, imp cap, rand%, net_n, ceiling, floor, consonance, Dissonance are **T**-theoretical elements.

(3) $y = \langle N, C, T, N^+, N^-, Pairs, Conpairs, W, net_n \rangle \in \mathbf{M_{pp}}$(HN for Consonant)

This modified version of the Hopfield network is a kind of specialization of the Hopfield network by adding those additional requirements above. Now modeling the intertheoretical connections between the Hopfield Network for consonance model **T**(HN for Consonance) and the Festinger Theory of Cognitive Dissonance **T**(DissB) can be done as follows. (1) The first step is defining the determining links, which describe the intertheoretical connections between the terms of both theories that are connected. (2) The second step is determining an echelon partial subset of the potential models of the theory of cognitive dissonance $\mathbf{M_p}$(DissB), whose elements are connected to the elements of the potential model of the Hopfield network for the consonance model. This echelon partial subset of $\mathbf{M_p}$(DissB) describes which parts of the dissonance theory can be reduced and simulated by the Hopfield network for the consonance model. The echelon partial subset of $\mathbf{M_p}$(HN for Consonance) will describe which parts of the Hopfield network for the consonance model reduce the dissonance theory (D VI-5). (3) We can determine the entailment link between the echelon partial subsets of both theories' potential models. (4) The last step will be projecting the intertheoretical determining links to the partial potential model of one of both theories to characterize the class of local empirical claims and the class of intended applications. The first three steps will be executed as follows:

DVI-5: Λ is a collection of determining links between **T**(DissB) and **T**(HN for Consonance), e_1 is an echelon partial subset of the potential models of the dissonance theory ($\mathbf{M_p}$(DissB)) and e_2 is an echelon partial subset of the potential models of the Hopfield network for the consonance model ($\mathbf{M_p}$(HN

for Consonance)), and E is a set of entailment links connecting both echelon partial subsets, e_1 and e_2, iff there exist x_1, x_2, λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, λ_7 such that:

(1) $x_1 = \langle$Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress$\rangle \in \mathbf{M_p}$(DissB)

(Let x_1 be a potential model of the cognitive dissonance ($\mathbf{M_p}$(DissB)).)

(2) $x_2 = \langle$N, C, T, N^+, N^-, Pairs, Conpairs, W, Resist, A, imp, cap, rand%, net_n, ceiling, floor, consonance, Dissonance$\rangle \in$ $\mathbf{M_p}$(HN for Consonance)

(Let x_2 be a potential model of the Hopfield network for the consonance model ($\mathbf{M_p}$(HN for Consonance)))

(3) $\lambda_1 \subseteq$ Cognition × Pairs

(λ_1 is the determining link that connects the set *Cognition* to the set *Pairs* and λ_1 is bijective)

(4) $\lambda_2 \subseteq$ Disscog × Conpairs

(λ_2 is the determining link that connects the set *Disscog* to the set *Conspairs* and λ_2 is bijective)

(5) $\lambda_3 \subseteq$ Conscog × Conpairs

(λ_3 is the determining link that connects the set *Conscog* to the set *Conspairs* and λ_3 is bijective)

(6) $\lambda_4 \subseteq$ pairdiss × W

(λ_4 is the determining link that connects the set *pairdiss* to the set of connection weight *W* and λ_4 is bijective)

(7) $\lambda_5 \subseteq$ paircons × W

(λ_5 is the determining link that connects the set *paircons* to the set of connection weight *W* and λ_5 is bijective)

(8) $\lambda_6 \subseteq$ Pairimp × imp

(λ_6 is the determining link that connects the set *Pairimp* to the set *imp* and λ_6 is bijective)

(9) $\lambda_7 \subseteq$ diss × Dissonance

(λ_7 is the determining link that connects the set *diss* to the set *Dissonance* and λ_7 is bijective)

(10) $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7\}$

(Λ is a collection of the determining links between ($\mathbf{M_p}$(DissB)) and ($\mathbf{M_p}$(HN for Consonance)))

(11) $e_1 = \langle$Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss$\rangle \in$ echelon partial subset of $\mathbf{M_p}$(DissB)

(e_1 is an echelon subset of ($\mathbf{M_p}$(DissB)))

(12) $e_2 = \langle$Pairs, Conpairs, W, Dissonance $\rangle \in$ echelon partial subset of $\mathbf{M_p}$(HN for Consonance)

(e_2 is an echelon subset of ($\mathbf{M_p}$(HN for Consonance)))

(13) $E = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7\}$

(E is the entailment link between the echelon subset of ($\mathbf{M_p}$(DissB)) and the echelon subset of ($\mathbf{M_p}$(HN for Consonance)))

The last step, which is to characterize the interpreting links on the partial potential models, aims to determine local empirical claims and the empirical contents of the intertheoretical reduction of cognitive dissonance theory by the Hopfield network according to the consonance model. Based on D VIII-7 and VIII-8 in BMS, pp. 398–400 the interpreting links for the reduction of the theory of cognitive dissonance **T**(DissB) by the Hopfield network for the consonance model **T**(HN for Consonance) can be determined as follows:

DVI-6: E*(DissB) = $\{l_1, l_2, l_3, l_4, l_5\, l_6, l_7\}$ is a collection of interpreting links, where **T**(DissB) is interpreted by **T**(HN for Consonance) in the consonance model and f_1 is the set of empirical claims of the interpreting links of the

reduction of DissB by HN for Consonance, iff there exist x_1, x_2 , e_1, e_2, E, y_1 such that:

(1) x_1 = ⟨Time, Rawcog, Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress⟩ ∈ $\mathbf{M_p}$(DissB)
(x_1 is the potential model of the theory of cognitive dissonance ($\mathbf{M_p}$ (DissB)))

(2) x_2 = ⟨N, C, T, N^+, N^-, Pairs, Conpairs, W, Resist, A, imp, cap, rand%, net_n, ceiling, floor, consonance, Dissonance⟩ ∈ $\mathbf{M_p}$(HN for Consonance)
(x_2 is the potential model of the Hopfield network for consonance ($\mathbf{M_p}$(HN for Consonance)))

(3) e_1 = ⟨Cognition, Disscog, Conscog, pairdiss, paircons, pairimp, diss⟩ ∈ echelon set of $\mathbf{M_p}$(DissB) connected to $\mathbf{M_p}$(HN for Consonance)
(e_1 is an echelon subset of ($\mathbf{M_p}$(DissB)))

(4) e_2 = ⟨ Pairs, Conpairs, W, imp, Dissonance ⟩ ∈ echelon set of $\mathbf{M_p}$(HN for Consonance) connected to $\mathbf{M_p}$(DissB)
(e_2 is an echelon subset of ($\mathbf{M_p}$(HN for Consonance)))

(5) E = {λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, λ_7} ∈ the set of entailment link between e_1 and e_2.
(E is the set of entailment links between DissB and HN for Consonance)

(6) y_1 = ⟨Time, Rawcog, Cognition, Disscog, Conscog⟩ ∈ $\mathbf{M_{pp}}$(DissB)
(y_1 is a partial potential model of the theory of cognitive dissonance ($\mathbf{M_{pp}}$ (DissB)))

(7) E*(DissB) = {l_1, l_2, l_3, l_4, l_5 l_6, l_7} ∈ the set of interpreting links, where {l_i = ⟨x', y⟩| x'∈$\mathbf{M_p}$(**T'**), y∈$\mathbf{M_{pp}}$*(**T**) and there is x∈$\mathbf{M_p}$(**T**) such that ⟨x',x⟩∈(**T'**, **T**) and r*(x)=y}.
(E*(DissB) is a collection of interpreting links)

(8) $l_1 = \{\langle pair_j , cog_i\rangle | \text{ Pair}(pair_j) \wedge \text{Cognition } (cog_i) \rightarrow R(pair_j , cog_i)\}$, where R(x,y) = x interprets y.

(l_1 is an interpreting link that interprets the concept of cognition represented by a pair of neurons. The set *Cognition* here is an element of $\mathbf{M_{pp}}$(DissF), whereas the set *Pairs* is an element of $\mathbf{M_p}$(HN for Consonance))

(9) $l_2 := \{\langle cp_j , dc_i\rangle | \text{ Conpair}(cp_j) \wedge \text{Disscog}(dc_i) \rightarrow R (cp_j, dc_i)\}$

(l_2 is an interpreting link that interprets the concept of dissonant cognitions being represented by some connection between a pair of neurons. The set *Disscog* here is an element of $\mathbf{M_{pp}}$(DissF), whereas the set *Conpair* is an element of $\mathbf{M_p}$(HN for Consonance))

(10) $l_3 = \{\langle cp_j, cc_i\rangle | \text{ Conpair}(cp_j) \wedge \text{Conscog}(cc_i) \rightarrow R (cp_j , cc_i)\}$

(l_3 is an interpreting link that interprets the concept of consonant cognitions being represented by some connection between a pair of neurons. The set *Conscog* here is an element of $\mathbf{M_{pp}}$(DissF), whereas the set *Conpair* is an element of $\mathbf{M_p}$(HN for Consonance))

(11) $l_4 = \{\langle w_j, pd_i\rangle | \ w_j \in [-1, 0[\wedge \text{pairdiss}(pd_i) \rightarrow R(w_j, pd_i)\}$

(l_4 is an interpreting link that interprets the concept of the dissonance within pairs of cognitions being represented by the synaptic weight. The set *pairdiss* here is an element of $\mathbf{M_{pp}}$(DissF), whereas the set *W* is an element of $\mathbf{M_p}$(HN for Consonance))

(12) $l_5 = \{\langle w_j, pc_i\rangle | \ w_j \in]0, 1] \wedge \text{paircons}(pc_i) \rightarrow R(w_j, pc_i)\}$

(l_5 is an interpreting link that interprets the concept of the consonance within pairs of cognitions being represented by the synaptic weight. The set *paircons* here is an element of $\mathbf{M_{pp}}$(DissF), whereas the set *W* is an element of $\mathbf{M_p}$(HN for Consonance))

(13) $l_6 = \{\langle imp_j, pi_i\rangle | \ imp_j \in \{0.5, 1, 1.5\} \wedge \text{pairimp}(pi_i) \rightarrow R(imp_j, pi_i)\}$

(l_6 is an interpreting link that interprets the concept of the importance within pairs of cognitions being represented by the synaptic weight. The set *Pairimp* here is an element of $\mathbf{M_{pp}}$(DissF), whereas the set *Imp* is an element of $\mathbf{M_p}$(HN for Consonance))

(14) $l_7 = \{\langle \text{dissonance}_j, \text{diss}_i\rangle | \text{ dissonance}(\text{dissonance}_j) \wedge \text{diss}(\text{diss}_i) \rightarrow R (\text{dissonance}_j, \text{diss}_i)\}$

(l_7 is an interpreting link that interprets the concept of the magnitude of dissonance being represented by the concept of dissonance. The set *diss* here is an element of $\mathbf{M_{pp}}$(DissF), whereas the set *dissonance* is an element of $\mathbf{M_p}$(HN for Consonance))

(15) $f_1 = \langle$Cognition, Disscog, Conscog$\rangle \in$ an echelon subset of $\mathbf{M_{pp}}$(DissB) by applying function r*: $e_1 \rightarrow f_1$, that mapping E* from $\mathbf{M_p}$ (DissB) to $\mathbf{M_{pp}}$(DissB)

(f_1 is the echelon subset of $\mathbf{M_{pp}}$(DissB) representing the local empirical claims of the intertheoretical reduction on the side of the reduced theory of cognitive dissonance (DissB))

Based on DVIII-9 and DVIII-10 in BMS, f_1 can be determined as the set of empirical claims of the intertheoretical reduction of the theory of cognitive dissonance to the Hopfield network for the consonance model. The empirical claims of this intertheoretical reduction refer to a set of the actual cognitions and the relations between these cognitions, either consonant or dissonant. The actual cognitions simulated are 'inputs' of the network, given by examples such as toy, punishment, playing or not playing, happy or not happy, etc. Every cognition is interpreted as a pair of neurons, that represent two poles of valuation. The pairs of cognitions, either dissonant or consonant, are interpreted as connected pairs of neurons in the Hopfield network for the consonance model.

A Summary of This Part. We have built a structuralist model to model and analyze the intertheoretical connections between the Festinger theory of cognitive dissonance (DissB) and the Hopfield network for the consonance model (HN for consonance). In this case only part of $\mathbf{M_p}$(DissB), namely the echelon partial subset c_1(DissB), is connected to part of $\mathbf{M_p}$(HN for consonance), i.e. the echelon partial subset c_2(HN for consonance). Besides modeling the intertheoretical connections on the theoretical level, the structuralist metatheory of science provides us also with the tools for modeling and analyzing the intertheoretical connections on the non-theoretical level i.e., the local empirical claims of the intertheoretical relation. Our model can be graphically represented as follows:

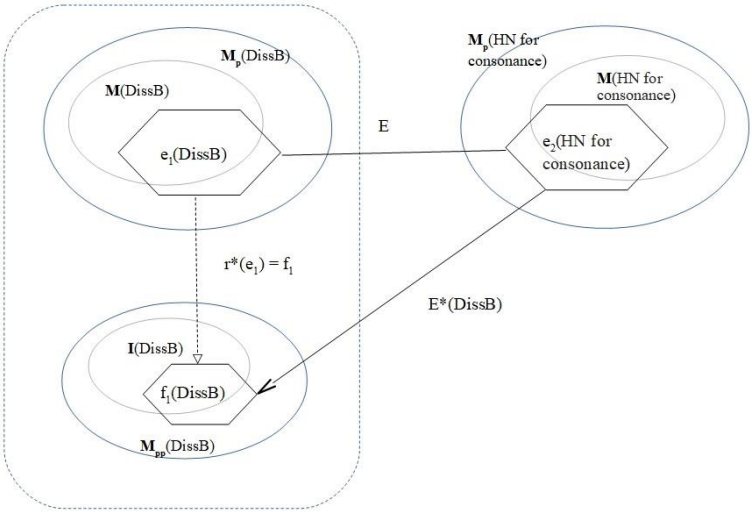


Figure 6.3. A structuralist modeling of intertheoretical reduction between the theory of cognitive dissonance (DissB) and the Hopfield network for the consonance model (HN for Consonance). e(DissB) is an echelon subset of $\mathbf{M_p}$(DissB) and e(HN for Consonance) is an echelon subset of $\mathbf{M_p}$(HN for Consonance). Both are connected by a set of entailment links (E). The interpreting links with local at DissB (E*(DissB)) connect the set of empirical claims of this intertheoretical reduction at DissB to the echelon subset of $\mathbf{M_p}$(HN for Consonant) as the reducing theory.

Chapter 7
The Structuralist Model of Intertheoretical Connections between the Festinger Theory of Cognitive Dissonance and the Two Layers Feed-Forward Neural Network in the Connectionist Model of Simulation

Another simulation of cognitive dissonance, whose intertheoretical connections will be modeled and analyzed here, is the connectionist model. This simulation implements the Rosenblatt perceptron with the two-layers feed-forward neural network and the delta rule as its learning rule. To model intertheoretical relations for this simulation that involves more than two theories for this simulation, the following strategy will be implemented: The first step is to model the intertheoretical connections that unify the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule. The result will be simplified by building a new theory-element that unifies the theory-elements of these three theories to make the modeling and analysis of intertheoretical connections for the simulation easier. The last step will be modeling and analysis of the intertheoretical connections of the simulation itself by using the unifying theory-element, instead of using the original theory-elements.

7.1. The Intertheoretical Connections Between the Two-Layers Feed-Forward Neural Network, the Rosenblatt Perceptron, and the Delta Rule

For building a model of the intertheoretical relation between these three theories, this dissertation will use the approach to model intertheoretical relations between two theory-elements. In the connectionist model, several perceptrons are placed in a two-layers feed-forward network,

and this network of neurons is trained by applying the delta rule. Based on this schema, two intertheoretical connections of two theory-elements will take part in modeling the intertheoretical connection between them, i.e., between the neurons and their network and between the network of neurons and the learning rule. In this intertheoretical relation, neither the potential models of the Rosenblatt perceptron ($\mathbf{M_p}$(RP)) are fully connected to the potential models of the two layers feed-forward neural networks ($\mathbf{M_p}$(2L-FFNN)) nor the potential models of the two layers feed-forward neural networks ($\mathbf{M_p}$(2L-FFNN)) to the potential models of the delta rule ($\mathbf{M_p}$(DR)). Hence, we should apply the definition of determining links to model and analyze them.

To reduce the complexity in discussing the idea behind this combination, let us first focus on the intertheoretical relation between the Rosenblatt perceptron and the two-layers feed-forward neural network. By placing several neurons in two-layers feed-forward architecture, the following concepts of both theory-elements are connected. The concept of neuron (N) in $\mathbf{M_p}$(RP) is connected with the concept of neuron (N) in $\mathbf{M_p}$(2L-FFNN). The concept of input neuron (N_0) in $\mathbf{M_p}$(RP) is also connected with the concept of input neuron (N_{in}) in $\mathbf{M_p}$(2L-FFNN). The concept of connections (C) in $\mathbf{M_p}$(RP) and the concept of connections (C) in $\mathbf{M_p}$(2L-FFNN) are connected. Moreover, the concept of synaptic weight (W) in the $\mathbf{M_p}$(RP) and the concept of synaptic weight (W) in the $\mathbf{M_p}$(2L-FFNN) are also connected. Finally, all three **T**-theoretical concepts of both $\mathbf{M_p}$(2L-FFNN) and $\mathbf{M_p}$(RP) are connected: *fnet* is connected with net_n, *fact* with act_n, and *fout* with out_n.

After placing those neurons in the network, the delta rule is applied as a learning rule for the network. The second step is to connect the concepts from the delta rule to the concepts of the network. The concept of neuron (N) from $\mathbf{M_p}$(DR) cannot be connected to the concept of neuron in the network because the delta rule is not applied to the input neurons.

Therefore, the concept of neuron from $\mathbf{M_p}$(DR) is not connected with all neurons in the input neurons (N_0 or N_{in}). In the case of the two layers feed-forward neural network, which does not have the hidden layer ($N_{hidden} = \varnothing$), the concept of neuron from $\mathbf{M_p}$(DR) is only connected to the concept of output neuron from the network (N_{out} and N/N_0). However, the concept of connection C from $\mathbf{M_p}$(DR) can safely be connected to the concept of connection in the network. The concept of bias (B) from $\mathbf{M_p}$(DR) can also be connected to the concept of bias (B) in the network. The concept of connection weight (W) from $\mathbf{M_p}$(DR) is connected not only with the concept of connection weight of the network but also with (W_0) from $\mathbf{M_p}$(RP). Weight (W) from $\mathbf{M_p}$(DR) combines both of them. The concept of input from $\mathbf{M_p}$(DR) is identical to the input of the network, especially because the network is a two-layers feed-forward neural network. The concept of actual output ($OUTn$) from $\mathbf{M_p}$(DR) can also be seen as identical with the concept of output in the network (out_n and $outp$). However, the concept of desired output (Out) cannot be connected to the term output of the network (directly) because it is the label of the training-sample. Now, the intertheoretical relations connecting the potential models of these theories can be defined as follows:

D VII-1 :If $\mathbf{T}$(RP) = ⟨$\mathbf{M_p}$(RP), $\mathbf{M}$(RP), $\mathbf{M_{pp}}$(RP), $\mathbf{I}$(RP)⟩ and $\mathbf{T}$(2L-FFNN) = ⟨$\mathbf{M_p}$(2L-FFNN), $\mathbf{M}$(2L-FFNN), $\mathbf{M_{pp}}$(2L-FFNN), $\mathbf{I}$(2L-FFNN)⟩ and $\mathbf{T}$(DR) = ⟨$\mathbf{M_p}$(DR), $\mathbf{M}$(DR), $\mathbf{M_{pp}}$(DR), $\mathbf{I}$(DR)⟩ then there exist determining links between $\mathbf{T}$(RP) and $\mathbf{T}$(ArchNN) and $\mathbf{T}$(DR) iff there exist x_1 and x_2 and x_3 such that:

(1) x_1 = ⟨N, N_0, IR, IN, C, W, B, W_0, Inp, Outp, fnet, fact, fout⟩ ∈ $\mathbf{M_p}$(RP) (x_1 is a potential model of the Rosenblatt perceptron ($\mathbf{M_p}$(RP)))

(2) x_2 = ⟨N', N_{in}, N_{out}, N_{hidden}, IR, C', pred(n), succ(n), W', ext_n, net_n, act_n, out_n⟩ ∈ $\mathbf{M_p}$(2L-FFNN)

(x_2 is a potential model of the two-layers feed-forward neural network ($\mathbf{M_p}$(2L-FFNN)))

(3) $x_3 = \langle N^*, IR, Inp^*, Out^*, C^*, L, B^*, W^*, OUTn, \eta, Error\rangle \in \mathbf{M_p}(DR)$

(x_3 is a potential model of the delta-rule ($\mathbf{M_p}$(DR)))

(4) $N \lambda N'$

(The concept of neuron (*N*) in the $\mathbf{M_p}$(RP) is connected with the concept of neuron (*N'*) in the $\mathbf{M_p}$(2L-FFNN), because both refer to all neurons in the networks)

(5) $N_{out} \lambda N^* \lambda N/N_0$

(The concept of neuron (*N**) in the $\mathbf{M_p}$(DR) is connected with the concept of neuron (*N*) in $\mathbf{M_p}$(RP) without referring to the input neurons (N_0) and the concept of output neuron (N_{out}) in $\mathbf{M_p}$(2L-FFNN))

(6) $N_0 \lambda N_{in}$

(The concepts of input neuron (N_0) in $\mathbf{M_p}$(RP) and (N_{in}) $\mathbf{M_p}$(2L-FFNN) are connected. The case would be different in the multi-layer feed-forward neural network)

(7) $C \lambda C' \lambda C^*$

(The concepts of connection (*C, C', C**) between neurons in all potential models are connected)

(8) $B \lambda B^*$

(The concepts of bias in both $\mathbf{M_p}$(RP) and $\mathbf{M_p}$(DR) are connected)

(9) $W \lambda W'$

(The concepts of connection weight in the potential models of the Rosenblatt perceptron and two-layers feed-forward neural networks are connected)

(10) $W^* \subseteq W \times W_0$

(The concept of connection weight (W*) in $\mathbf{M_p}$(DR) is connected with both the concept of connection weight (W) and the concept of connection weight of the bias (W_0) in $\mathbf{M_p}$(RP))

(11) Inp λ Inp* λ ext_n

(The concept of input (Inp) in $\mathbf{M_p}$(RP) is connected with both the concept of external input (ext_n) in $\mathbf{M_p}$(2L-FFNN) and the concept of input (Inp*) in $\mathbf{M_p}$(DR))

(12) Outp λ out_n λ OUTn

(The concept of Output ($Outp$) in $\mathbf{M_p}$(RP) is connected with the concept of output (out_n) in $\mathbf{M_p}$(2L-FFNN) and the concept of actual output (OUT_n) in $\mathbf{M_p}$(DR))

(13) fnet λ net_n

(The concept of *fnet* in $\mathbf{M_p}$(RP) is connected with the concept of net_n in $\mathbf{M_p}$(2L-FFNN))

(14) fact λ act_n

(The concept of *fact* in $\mathbf{M_p}$(RP) is connected with the concept of act_n in $\mathbf{M_p}$(2L-FFNN))

Based on these connections, the intertheoretical connection connecting the actual models of these theories can be specified as follows:

D VII-2: If $\mathbf{T}$(RP) = ⟨$\mathbf{M_p}$(RP), $\mathbf{M}$(RP), $\mathbf{M_{pp}}$(RP), $\mathbf{I}$(RP)⟩ and $\mathbf{T}$(2L-FFNN) = ⟨$\mathbf{M_p}$(2L-FFNN), $\mathbf{M}$(2L-FFNN), $\mathbf{M_{pp}}$(2L-FFNN), $\mathbf{I}$(2L-FFNN)⟩ and $\mathbf{T}$(DR) = ⟨$\mathbf{M_p}$(DR), $\mathbf{M}$(DR), $\mathbf{M_{pp}}$(DR), $\mathbf{I}$(DR)⟩ then there exist determining links among $\mathbf{T}$(RP) and $\mathbf{T}$(ArchNN) and $\mathbf{T}$(DR) iff there exist x_1 and x_2 and x_3 such that:

(1) x_1 = ⟨N, N_0, IR, IN, C, W, B, W_0, Inp, Outp, fnet, fact, fout⟩ ∈ $\mathbf{M_p}$(RP)

(x_1 is a potential model of the Rosenblatt Perceptron ($\mathbf{M_p}$(RP)))

(2) x_2 = ⟨N', IR, N_{in}, N_{out}, C', pred(n), succ(n), W', ext_n, net_n, act_n, out_n⟩ ∈ $\mathbf{M_p}$(2L-FFNN)

(x_2 is a potential model of the two-layers feed-forward neural network ($\mathbf{M_p}$(2L-FFNN)))

(3) x_3 = ⟨N*, IR, Inp*, Out*, C*, L, B*, W*, OUTn, η, Error⟩ ∈ $\mathbf{M_p}$(DR)

(x_3 is a potential model of the delta rule ($\mathbf{M_p}$(DR)))

(4) N = N'

(The concept of neuron (*N*) of the Rosenblatt perceptron is identical to the concept of neuron (*N'*) of the two-layers feed-forward neural network)

(5) N_{out} = N* = N/N_0

(The concept of neuron (*N**) of the delta rule refers to the set of neurons (*N*) in the perceptron without referring to the set of neurons input (N_0). Moreover, in the two-layers feed-forward neural network this only refers to the neuron output (N_{out}))

(6) $N_0 = N_{in} \rightarrow$ Inp

(The concept of neuron input (N_0) of the Rosenblatt perceptron is identical to the concept of neuron in the input layer (N_{in}) of the two-layers feed-forward neural network. In applying the delta rule, real numbers are attached to these neurons as the input training for the network)

(7) C = C' = C*

(The concept of connection in all theory-elements is identical)

(8) B = B*

(The concept of bias in the Rosenblatt perceptron and the concept of bias in the delta rule are identical)

(9) W = W'

(The concept of synaptic weight in the Rosenblatt perceptron and the concept of the two layers feed-forward neural network are identical)

(10) $W^* := W_0 \cup W$

(The concept of synaptic weight W^* in the delta rule refers to unification between W_0 and W in the Rosenblatt perceptron)

(11) $Inp = Inp^* = ext_n$

(The concept of input from the delta rule is identical to the concept of input of the Rosenblatt perceptron)

(12) $Outp = Out_n = OUTn$

(The actual output of the Rosenblatt perceptron is identical with the concept of actual output of the delta rule)

There is $n \in N/N_0$, $c_i \in C$ for $i \in IN$, $b \in B$ and let net_n, act_n, out_n so that:

(13) $net_n = fnet\ (Inp, W) = \sum_{i=1}^{n} Inp_i\ (n, c_i,) \cdot W(c_i)$

(For each output neuron: its network input (net_n) is the result of the network input function of the neuron. The neurons receive input according to $net_n = fnet\ (Inp, W) = \sum_{i=1}^{n} Inp_i\ (n_i, c_i) \cdot W(c_i)$)

(14) $act_n = fact\ (net_n, b, w_0)$ according to linear regression: $fact\ (net_n, b, w_0) = net_n + b \cdot w_0$

(For each output neuron: Its activation (act_n) is the result of the activation function of the neuron, and its network output is the result of the output function of the neurons. The neurons are activated (in an excitatory state) according to the linear regression: $fact(net_n, b, w_0) = net_n + b \cdot w_0$)

(15) $out_n = fout\ (act_n) = Outp$

(For each output neuron: it sends its output signal according to fout(fact), which is identical to the output (out_n) in the two layers feed-forward neural network and the actual output ($Outp$) in the delta rule)

From the formulation of the intertheoretical connections above, there are several concepts in the three theories that are identical, especially in the case of the two-layers feed-forward neural network. Therefore, the model's simplification in a unifying model can be done by omitting the models' redundant concepts as follows:

(1) Because the set N in the $\mathbf{M_p}$(RP) and N' in the $\mathbf{M_p}$(2L-FFNN) are identical, we will use just N for our new unifying models.

(2) Because the set N_0 in the $\mathbf{M_p}$(RP) and N_{in} in the $\mathbf{M_p}$(2L-FFNN) are identical, we will use just N_{in} for our new unifying models.

(3) Because the set N^* in the $\mathbf{M_p}$(DR) and N/N_0 in the $\mathbf{M_p}$(RP) refer to the same set of N_{out} in $\mathbf{M_p}$(2L-FFNN) in the new unifying model, we will use just N_{out}.

(4) We will use only the set C because all the connected terms come from the three theories.

(5) We will use just the set of bias B from the Rosenblatt Perceptron because this set is identical with the set of bias B^* in the delta rule.

(6) We will use the sets W and W_0 in our new unifying models because both terms can cover both sets W' and W^*.

(7) In the case of the two layers feed-forward neural networks, all input neurons are connected to every output neuron. Therefore, all sets of inputs (*Inp, extn, Inp**) from all models refer to the same inputs. Therefore, we use only the term *Inp* for our new models.

(8) In the case of the two layers feed-forward neural networks, the outputs of every neuron besides the input neuron (*Outp*) in the Rosenblatt perceptron are the outputs (out_n) of the neural networks and are also seen as the actual output (*OUTn*) according to the delta rule. Therefore, we will use just the set out_n in our new unifying models.

(9) We will also take over *pred(n), succ(n), netn, actn* from the two layers feed-forward neural networks. From the Rosenblatt Perceptron, we need three functions, namely *fnet, fact,* and *fout*. Moreover, from the delta rule, we need the set of desired outputs (*Out*), The set of the training-sample (*L*), the learning rate (η) and the set *Error* from the delta rule.

(10) We will still need to add the set of natural numbers and the set of rational numbers.

With these considerations, the standard form of the unifying potential model of the Rosenblatt perceptron, the two layers feed-forward neural network, and the delta rule ($\mathbf{M_p}$(RP + 2L-FFNN + DR)) can be defined as follows:

D VII-3: x is a unifying potential model of the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule ($\mathbf{M_p}$(RP + 2L-FFNN + DR)) iff there exist N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, η, fnet, fact, fout, net_n, act_n, out_n, Error such that:

(1) $x = \langle N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, \eta, fnet, fact, fout, net_n, act_n, out_n, Error\rangle \in \mathbf{M_p}$(RP + 2L-FFNN + DR)

(2) N = a finite non-empty set of neurons.

(3) N_{in} = non-empty set of input neurons.

(4) N_{out} is a finite non-empty set of output-neurons

(5) $C \subseteq N \times N$

(a finite non-empty set of connection between neurons)

(6) $pred(n) = \{n_1 \in N | (n_1, n_2) \in C\}$ (presynaptic neurons)

(7) $succ(n) = \{n_2 \in N | (n_1, n_2) \in C\}$ (postsynaptic neurons)

(8) $W := C \rightarrow IR$

(Synaptic Weight *W* assigns to each pair of neurons a real number as synaptic weight.)

(9) $B := N \rightarrow IR$

(Bias *B* assigns to every neuron a real number as its bias. Bias is normally set = 1)

(10) $W_0 := B \times N \rightarrow IR$ (synaptic weight from bias)

(11) $Inp := N_{out} \times C \rightarrow IR$

(Input *Inp* assigns to each neuron several real numbers as its input, that is sent by its Input-units (N_0) in the network)

(12) $fnet := W \times Inp \rightarrow IR$

(Network Input function (*fnet*) assigns to Neuron (except for input unit) a real number as network input)

(13) $fact := fnet \times b \rightarrow IR$ (activation function)

(14) $fout := fact \rightarrow Outp$ (output function)

(15) $ext_n := N_{in} \rightarrow IR$ (external input)

(16) $net_n := N_{out} \rightarrow IR$ (network-input)

(17) $act_n := N_{out} \rightarrow IR$ (activation)

(18) $out_n := N_{out} \rightarrow IR$ (output)

(19) $Inp \subseteq IR$

(input – $\overrightarrow{Inp}$ is a set of the input-vector, whose elements are identical with *Inp*)

(20) $Out \subseteq IR$

(desired output – $\overrightarrow{Out}$ is a set of an output-vector, whose elements are identical with *Out*)

(21) $L \subseteq Inp \times Out$

(a finite non-empty set of training examples)

(22) $Out_n \subseteq IR$

(actual output, if the neural network is fed with input *Inp*. $\overrightarrow{Out_n}$ is a set of an output-vector, whose elements are identical with Out_n)

(23) $\eta \in IR$ (learning rate)

(24) $Error := Out \times Out_n \rightarrow IR^2$

(The network's error mapping in a two-dimensional Cartesian coordinate system)

The unifying actual models of the new (synthetized) theory can be characterized as follows:

D VII-4 :x is a unifying actual model of the Rosenblatt perceptron, the two layers feed-forward neural network, and the delta rule (**M**(RP + 2L-FFNN + DR)) iff there exists N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, η, fnet, fact, fout, net_n, act_n, out_n, Error such that:

(1) $x = \langle N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, \eta, fnet, fact, fout, net_n, act_n, out_n, Error \rangle \in \mathbf{M_p}$(RP + 2L-FFNN + DR)

(2) $N = N_{in} \cup N_{out}$

(3) for all $n \in N$ it holds:

(3.1) $N_{in}(n) \leftrightarrow pred(n)$

(3.2) $N_{out}(n) \leftrightarrow succ(n)$

(4) $N_{in} \cap N_{out} = \varnothing$

(5) $C \subseteq N_{in} \times N_{out}$

(6) There is $n \in N_{out}$, $c_i \in C$ for $i \in IN$, $b \in B$ and let net_n, act_n, out_n so that:

(6.1) $net_n = fnet\ (Inp, W) = \sum_{i=1}^{n} Inp_i\ (n, c_i,).W(c_i)$

(6.2) $act_n = fact(net_n, b, w_0)$ according to Linear Regression: $fact(net_n, b, w_0) = net_n + b.w_0$

(6.3) $out_n = fout\ (act_n)$.

(7) $\forall i \in \{1,...,n\}$: $Error = \frac{1}{2} \Sigma_i\ (Out_n - Out)^2$ and the derivation for each neuron's activation: $(Out_n - Out)$

(8) $\forall b_i \in B$, $i=1,...,n$: $b_i^{(new)} = b_i^{(old)} + \Delta b_i$ with $\Delta b_i = -\eta(Out_n - Out)$.

(9) $\forall w_i \in W$, i=1,...,n: $w_i^{(new)} = w_i^{(old)} + \Delta w_i$ with $\Delta w_i = -\eta(Out_n - Out)\ Inp_i$

(10) Convergence-statement:

Suppose $L=\{(Inp_1^{\rightarrow}, Out_1),..., (Inp_m^{\rightarrow}, Out_m)\}$ is a set of training-sample with

$L_0 = \{(Inp^{\rightarrow}, Out) \in L | Out = 0\}$ and $L_1 = \{(Inp^{\rightarrow}, Out) \in L | Out = 1\}$.

If L_0 and L_1 are linearly separable, viz. if $w^{\rightarrow} \in IR^n$ and $\theta = IR$ exist, so that:

$\forall (Inp^{\rightarrow}, 0) \in L_0 : w^{\rightarrow}\ Inp^{\rightarrow} < \theta$ and

$\forall (Inp^{\rightarrow}, 1) \in L_1 : w^{\rightarrow}\ Inp^{\rightarrow} \geq \theta$.

From the unifying potential models, the unifying partial potential models can be defined by omitting the **T**-theoretical elements (see Chapter 3) as follows:

D VII-5: y is a unifying partial potential model of the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule ($\mathbf{M_{pp}}$(RP + 2L-FFNN + DR)) iff there exist x such that:

(1) $x = \langle N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, \eta, fnet, fact, fout, net_n, act_n, out_n, Error\rangle \in \mathbf{M_p}$(RP + 2L-FFNN + DR)

(2) fnet, fact, fout, net_n, act_n, out_n, η, Error are **T**-theoretical.

(3) $y = \langle N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L\rangle \in \mathbf{M_{pp}}$(RP + 2L-FFNN + DR)

The relation between the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule is a relationship between three theories at the same time, even if the links being used always exist only between two theories. This “tripartite” relation occurs if we combine two (or maybe more) theory-elements ($\mathbf{T_1, ..., T_n}$) into another theory-element $\mathbf{T_0}$,

which serves as the mainboard theory – because of its form I call this relation the *"V-pattern of intertheoretical relations."* The notion of the V-pattern and the unifying theory-element will be discussed in more detail in Chapter 8. This relation can be described in a directed acyclic graph – like the letter "V" – as follows:

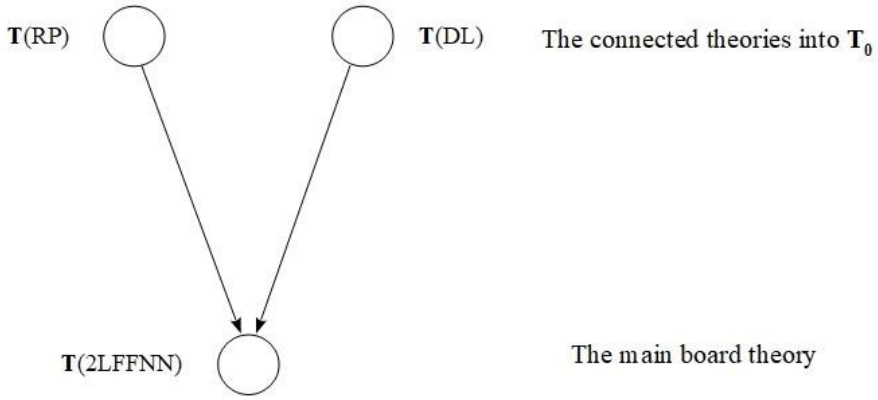


Figure 7.1. The directed acyclic graph of the V-pattern of intertheoretical relation. In our case, the mainboard theory is the two-layers feed-forward neural network, a specialization of the architecture of the neural network. The connected theories are the perceptron and the delta-rule. The word "mainboard theory" are used by referring to the fact that the other combined theories are connected to this theory and the local intended applications of intertheoretical connections of this combination of theories lay on it.

7.2. The Connectionist Model

The connectionist model was created by Frank van Overwalle and Karen Jordens, as a further advance in modeling cognitive dissonance. This model deals with some aspects that are not covered by the consonance model. "In this [consonance] model cognitions about discrepant behavior, justification, and evaluation are represented in separate nodes, and

connection weights denote the causal implications between cognitions, much like in automatic spreading activation models. Shultz and Lepper's novel contribution is that the consonance model can reach consistency automatically through the simultaneous satisfaction of the multiple constraints imposed by the connections. However, an important limitation is that the connections themselves are not dynamically learned, but handset by the authors based on available evidence. The aim of this article is to further advance the connectionist modeling of cognitive dissonance by presenting an alternative connectionist model in which the connections between cognitions are automatically developed, without intervention from the experimenter" (Overwalle and Jordens, 2002, p. 204). The connectionist model's aims are very different from the consonance model, and its set of basic assumptions on how the mind works are very different as well. "... our connectionist approach reflects a view of the mind as an adaptive learning mechanism, where cognitive dissonance is seen as a relatively rational process in which people seek causal answers for why they think, feel or behave inconsistently" (van Overwalle and Jordens, 2002, p. 205).

The basic idea of this simulation is inspired by the attributional reformulation of the Festinger theory of cognitive dissonance advocated by Cooper and Fazio (1984): "Cognitive dissonance reduction is driven by a rational process in which the causal understanding of thoughts, feelings, and behaviors plays a major role" (van Overwalle and Jordens, 2002, p. 205). This idea agrees with Cooper and Fazio's notion that people's attempt to understand and justify their dissonant behavior and emotions causally is at the root of the creation and reduction of dissonance (van Overwalle and Jordens, 2002, p. 205). However, there are also several differences between both as follows. (1) In Cooper and Fazio's notion one's responsibility takes a central role in changing one's attitudes to justify their discrepant behavior in the dissonance situation, whereas in the connectionist model, it is the attitude object that takes the central role. "We view the attributions to the

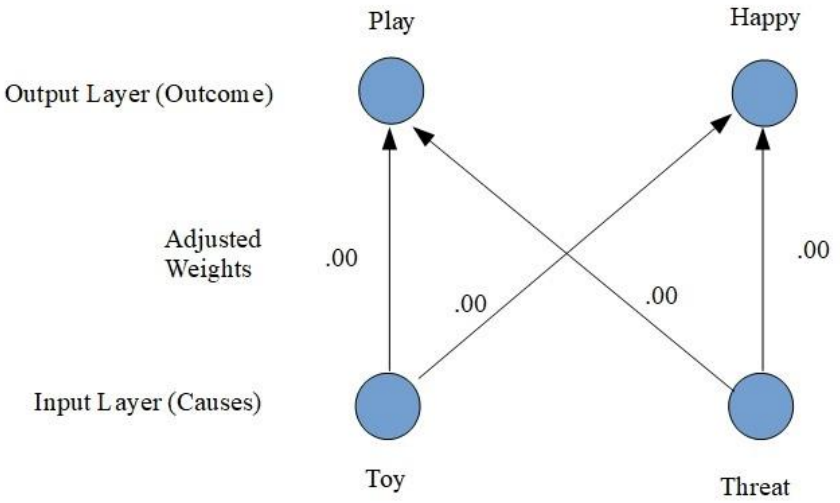
attitude object as central rather than attributions of the one's responsibility, we emphasize the role of affect during dissonance and neglect arousal, and we focus on unexpected outcomes rather than unwanted outcomes" (van Overwalle and Jordens, 2002, p. 205). To specify their attitude objects, Overwalle and Jordens follow the three components view on attitude from Rosenberg and Hovland (1960): cognitions, behaviors, and emotions. They define "an attitude as manifesting itself through its causal connections in memory between the cognitive representation or belief about the attitude object and feelings about this interaction. The intensity of an attitude is defined by the strength of these connections" (van Overwalle and Jordens, 2002, p. 205). (2) In this connectionist model, the dissonance is an emotional state of discomfort rather than physiological arousal. This kind of emotion serves as a source of information in making judgments and inferences in a dissonance situation. Moreover, affection experience itself is subjected to an attributional analysis. Also, in Cooper and Fazio's model, dissonance creates arousal, which serves as the instigator of an attributional interpretation (van Overwalle and Jordens, 2002, p. 206).

This model implements the two-layers feed-forward neural networks and the delta rule as its training algorithm (van Overwalle and Jordens, 2002, pp. 206–207) to simulate a specific example of cognition experiment called the first insufficient justification paradigm by Freedman (1965).[1] This

[1] The experiment is conducted as follows: "School children were forbidden to play with an attractive toy (a robot) under either mild or severe threat of punishment, and the experimenter either stayed in the room while the child played (surveillance condition), or went away (this surveillance variable was not included in the introductory example). Actual play with the previously forbidden toy about 40 days later in the absence of the experimenter or any threat, revealed greater derogation of the forbidden toy in the mild than in the severe threat condition when there had been no surveillance. When there had been surveillance, the effect of severity of threat was negligible. … The attributional explanation for these results was that mild threat alone provided insufficient justification for the counter attitudinal behavior of not playing with the attractive toy and thus created high dissonance that was reduced by lowering the attraction for the toy. In contrast, either the high threat or the exper imenter's surveillance provided sufficient justification for not playing with the toy and thus created little dissonance and little attitude change" (van Overwalle and Jordens, 2002, p. 216).

neural network links the causes with the outcomes in its connections. In the connectionist model, an attitude object (such as attractive toys) and several additional external pressures (such as the mild or severe threat of punishment) are the causes, whereas a behavior (such as play) and a current emotion (such as happy) are the outcomes. The nodes representing causes and outcomes are located in two different layers that are connected via adjustable connections. The first layer contains the input nodes representing the possible causes, and the second layer contains the output nodes representing the outcome. The connections between both layers represent causal explanations. The adjustable connection weights represent the quality or strength of the causal influence, and for attitude-object nodes, they represent the intensity of the attitude. Activation in the network spreads from the input nodes to the output nodes through these connections. We take the example given by the authors to make it clear (Figure 2):

A. Feed-forward Network (initials weights)



B. Feed-forward Network (after prior history)

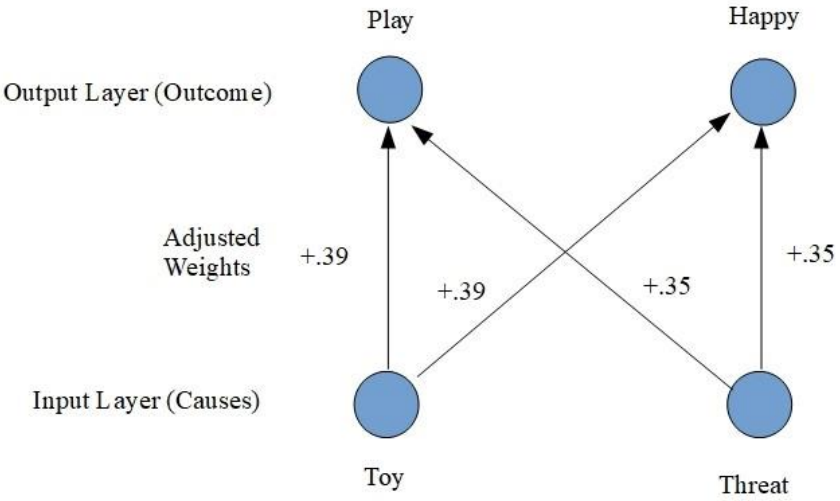


Figure 7.2. Specification of the feed-forward network model
(Source: Van Overwalle and Jordens, 2002, p. 207)

The delta rule strives to reduce the error between what the network expects from prior information and the current information. In the beginning, all connection weights are set at zero and eventually reach

excitatory, inhibitory, or zero weight depending on the person's learning history. In general, the delta rule predicts that the more a cause and an outcome co-occur, the stronger their connections will become until they reach the asymptote (typically -1 and +1). This learning process can be explained as follows (by using Figure 7.2): Initially, all weights are set to zero values (see Figure 7.2 A). Now we set a causal factor, such as a toy +1 (means present). This input will spread in proportion to the connection weight to all output nodes. It is because, in this initial condition, all connection weights are zero, and the activation weight of all output nodes is zero. The network here uses linear activations. "The activations received at the output nodes are linearly summed to determine their activation. This output activation can be understood as representing the magnitude of the outcome anticipated by the network" (van Overwalle and Jordens, 2002, pp. 207–208).

The observed outcome is represented by an external teaching signal, which has activation of +1 when the outcome is present (e.g., play or happy), zero when absent (e.g., not play, or moderate affect), and -1 when the opposite outcome is present (e.g., unhappy). The predicted outcome (output activation) is then compared with the actual occurrence of the outcomes (external teaching signal). In the beginning, output activation is zero, whereas now the actual output is +1, represented by playing with a toy and being happy. Thus, there is an error +1 for each output node. It means "the network at this point seriously underestimates the magnitude of the behavioral and emotional reactions" (van Overwalle and Jordens, 2002, p. 208).

The delta rule plays an important role here to make this simulation realistic. By implementing the delta rule, the network's connection weights are adjusted to minimize the discrepancy between predicted and actual output in proportion to the magnitude of the error. In this case, the connection between the toy and the outcomes will be adjusted upward. How

the delta rule works has already been explained in Chapter 3. We now introduce the threat, either +1 for severe threat or +0.5 for mild threat. The same process will happen, namely: (1) The neural network gets a new predicted outcome. (2) This predicted outcome will be compared with the new observed outcome, and the network gets a new discrepancy between both outcomes. (3) The delta rule will adjust the connection weights of the network according to the new situation. The condition of the network after certain trials is described in Figure 7.2 B.

How fast a person's mental representation of a dissonant situation is brought into correspondence with reality is represented by a learning rate parameter in the delta rule. This learning rate usually is between zero and +1. "A high learning rate indicates that new information has strong priority over old information and leads to radical adjustments in the connection weights, whereas a low learning rate suggests conservative adjustments that preserve much of the knowledge in the weights acquired by the old information" (van Overwalle and Jordens, 2002, p. 208).

This model reflects the dissonance reduction as follows: "… the discrepancy in the expected and actual outcomes (actions and affect) reflects cognitive dissonance, while the adjustments in the connection weights (determined by the delta algorithm) reflect dissonance reduction through attitude change" (van Overwalle and Jordens, 2002, p. 208). This simulation refers to Festinger's idea that "behavior is guided by accurate information about the environment and the self, and that dissonance can arise when this information disconfirms cognitions or expectations …. Therefore, any discrepancy between one's predictions (based on relevant input) and one's behavior or emotion would be disturbing to the person and will be avoided" (van Overwalle and Jordens, 2002, p. 209). The connectionist model captures dissonance by the fact that activation of the causal nodes will always create an error at the output without attitude adjustments. The magnitude of this error represents the magnitude of dissonance.

The simulation is conducted in two phases. In the first phase, called the pre-experimental phase, the connection weights are set in such a way as to simulate the set of beliefs and evaluations that the participants have. In the second phase, the experimental manipulation was closely replicated. "We first describe, how often the attitude object and external factors in the simulations occurred and under which experimental conditions, next the nature and direction of the behavioral and affective outcomes, how all these cognitions were coded in a distributed manner, and we end with some general features of the simulation. Although some of the specifications detailed next may seem arbitrary, they are in fact irrelevant with respect to the basic mechanism at work, and many of them can be relaxed without affecting the simulation result much (robustness section at the end of the simulations). We aim to demonstrate that some plausible assumptions about learning histories can explain human dissonance data, not that the specifications are necessarily correct nor that they are the only possible ones that make the simulation work" (Overwalle and Jordens, 2002, p. 213).

7.3. Festinger's Theory of Cognitive Dissonance and the Feed-Forward Neural Network for the Connectionist Model

The last intertheoretical relation that will be examined in this dissertation is the relation between the theory of cognitive dissonance and the implementation of the two-layers feed-forward neural network to build the connectionist model or simulation of dissonance reduction. For analysis the intertheoretical connections between the theory of cognitive dissonance and the feed-forward neural network according to the connectionist model, several preparatory steps are needed. The first step is adaptation of the unifying model of the perceptron, the two-layers feed-forward neural network, and the delta rule above to meet the requirements of the connectionist model. The second is the modification of the structuralist model of the theory of forced compliance dissonance (**T**(DissF)) from

Chapter 3 to fit the simulation's goal. (3) The last step is modeling the intertheoretical relation between both theories according to the connectionist model.

7.3.1. Adapting the Unified Model of the Feed-forward Neural Network to the Requirements of the Connectionist Model

From the description of the connectionist model above, the requirements can be summarized as follows: (1) The connectionist model uses the two-layers feed-forward neural network with the delta rule as its learning-rule. The neural networks consist of four neurons divided into two layers; each layer consists of two neurons. Two neurons in the input layer represent, for example, cognition of toy and cognition of threat, whereas the two neurons in the output layer represent cognition of play and cognition of happy. (2) The connectionist model implements the linear activations for the output neurons in the neural network. (3) The activation of an input-neuron is set 1 or 0 for a toy, which represents "present" or "not present." For the threat, an input neuron will be set 0, +0.5, or +1 to represent "not present," "mild threat," or "severe threat." (4) The expected activation of output neurons will be set at +1 when the outcome is present (e.g., play or happy), zero when absent (e.g., not play, or moderate affect), and -1 when the opposite outcome is present (e.g., unhappy). (5) And the learning rate for this simulation is set between 0 and 1. Based on these requirements, the unified theory element **T**(RP + 2L-FFNN + DR for Connectionist) should be adjusted by the following steps: First, we specialize the unifying potential models by adding or adjusting the specific required conditions as follows:

DVII-6: x is a unifying potential model of the Rosenblatt Perceptron, the two layers feed-forward neural network, and the delta rule for the connectionist model ($\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist)) iff there

exist N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, η, fnet, fact, fout, net_n, act_n, out_n, Error such that:

(1) $x = \langle N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, \eta, fnet, fact, fout, net_n, act_n, out_n, Error \rangle \in \mathbf{M_p}$(RP + 2L-FFNN + DR)

(2) N = a finite non-empty set of neurons

(3) N_{in} = non-empty set of input neuron

(4) N_{out} is a finite non-empty set of output neurons

(5) $C \subseteq N \times N$

(a finite non-empty set of connection between neurons)

(6) $pred(n) = \{n_1 \in N | (n_1, n_2) \in C\}$ (presynaptic neurons)

(7) $succ(n) = \{n_2 \in N | (n_1, n_2) \in C\}$ (postsynaptic neurons)

(8) $W := C \rightarrow IR$

(synaptic weight *W* assigns to each pair of neurons a real number as synaptic weight.)

(9) $B := N \rightarrow IR$

(bias – assigns to every neuron a real number as its bias. Bias is normally set = 1)

(10) $W_0 := B \times N \rightarrow IR$ (synaptic weight from bias)

(11) $Inp := N_{out} \times C \rightarrow IR$

(input – assigns to each neuron several real numbers as its input, that is sent by its input neurons N_0 in the network)

(12) $fnet := W \times Inp \rightarrow IR$

(network input function – assigns to Neuron except for input neurons a real number as network input)

(13) $fact := fnet \times b \rightarrow IR$

(activation function – there are various activation function)

(14) $fout := fact \rightarrow Outp$

(output function – assigns every neuron a real number as its output according to Outp)

(15) $ext_n := N_{in} \rightarrow IR$ (external input)

(15.1) $net_n := N_{out} \rightarrow IR$ (network-input)

(15.2) $act_n := N_{out} \rightarrow IR$ (activation)

(15.3) $out_n := N_{out} \rightarrow IR$ (output)

(16) $Inp \subseteq \{0, 0.5, 1\}$

(input – $Inp^{\rightarrow}$ is a set of the input-vector, whose elements are identical with Inp)

(17) $Out \subseteq \{-1, 0, 1\}$

(desired output – $Out^{\rightarrow}$ is a set of an output-vector, whose elements are identical with Out)

(18) $L \subseteq Inp \times Out$ (a finite non-empty set of Training example)

(19) $Out_n \subseteq IR$

(actual output, if the neural network is fed with input Inp. $Out_n^{\rightarrow}$ is a set of an output-vector, whose elements are identical with Out_n)

(20) $\eta \in]0, 1[$ (learning rate)

(21) $Error := Out \ni OUT_n \rightarrow IR^2$

(The Network's Error mapping in a two-dimensional Cartesian coordinate system)

The unifying actual models can be adjusted as follows:

D VII-7: x is a unifying actual model of the Rosenblatt perceptron, the two layers feed-forward neural network, and the delta rule for the connectionist model (**M**(RP + 2L-FFNN + DR for Connectionist)) iff there exist N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, η, fnet, fact, fout, net_n, act_n, out_n, Error such that:

(1) $x = \langle N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, \eta, fnet, fact, fout, net_n, act_n, out_n, Error \rangle \in \mathbf{M_p}$(RP + 2L-FFNN + Dr for Connectionist)

(Let x be a potential model of the theory unifying the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule)

(2) $N = N_{in} \cup N_{out}$

(All neurons are categorized in input neurons or output neurons)

(3) for all $n \in N$ it holds:

(3.1) $N_{in}(n) \leftrightarrow pred(n)$

(input neurons are neurons followed by other neurons in the network. They send their output to other neurons)

(3.2) $N_{out}(n) \leftrightarrow succ(n)$

(output neurons are neurons following other neurons in the network. They receive their input from other neurons)

(4) $N_{in} \cap N_{out} = \varnothing$

(There are no neuron playing both roles. There are only either input neurons or output neurons)

(5) $C \subseteq N_{in} \times N_{out}$

(All network connections are between input neurons and output neurons)

(6) For all $n_j \in N_{out}$, every $n_i \in N_{in}$, $c_{ij} \in C$ for $i,j \in IN$, $b \in B$ so that:

(6.1) $net_n = fnet(Inp, W) = \sum_{i=1}^{n} Inp_i(n_i, c_i,) \cdot W(c_i)$

(network input of output neurons)

(6.2) $act_n = fact(net_n, b, w_0)$ according to $fact(net_n, b, w_0) = net_n + b \cdot w_0$

(activation function of output neurons)

(6.3) $out_n = fout(act_n)$

(output of the output neurons)

(7) $\forall i \in \{1,...,n\}$: Error= ½ Σ_i $(Out_n - Out)^2$ and the derivation for each neuron's activation: $(Out_n - Out)$ (network error)

(8) $\forall b_i \in B$, i=1,...,n: $b_i^{(new)} = b_i^{(old)} + \Delta b_i$ with $\Delta b_i = -\eta(Out_n - Out)$ (updating bias)

(9) $\forall w_i \in W$, i=1,...,n: $w_i^{(new)} = w_i^{(old)} + \Delta w_i$ with $\Delta w_i = \eta(Out_n - Out)\ Inp_i$ (updating connection weights)

(10) Convergence-statement:

Suppose $L=\{(\vec{Inp_1}, Out_1),..., (\vec{Inp_m}, Out_m)\}$ is a set of training-sample with

$L_0 = \{(\vec{Inp}, Out) \in L | Out = 0\}$ and $L_1 = \{(\vec{Inp}, Out) \in L | Out = 1\}$.

If L_0 and L_1 are linearly separable, viz. if $\vec{w} \in IR^n$ and $\theta \in IR$ exist so that

$\forall(\vec{Inp}, 0) \in L_0$: $\vec{w}\ \vec{Inp} < \theta$ and

$\forall(\vec{Inp}, 1) \in L_1$: $\vec{w}\ \vec{Inp} \geq \theta$.

In the final step, the united partial potential models are characterized by omitting the **T**-theoretical elements (see Chapter 3) from the unifying potential models as follows:

DVII-8: y is a unified partial potential model of the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule for the connectionist model ($\mathbf{M_{pp}}$(RP + 2L-FFNN + DR for Connectionist)) iff there exist x such that:

(1) $x = \langle N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, \eta, fnet, fact, fout, net_n, act_n, out_n, Error\rangle \in \mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist)

(2) fnet, fact, fout, net_n, act_n, out_n, η, Error are **T**-theoretical.

(3) $y = \langle N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L\rangle \in \mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist)

7.3.2. Modification of the Theory-Element of Forced Compliance Dissonance for the Connectionist Model

The connectionist model attempts to simulate a single case; namely, the dissonance arouses by a given punishment if there is a toy to play. This case is a case of forced compliance dissonance (DissF), where the compliance is given in the form of punishment. We need to make some small modifications to the theory-element of the forced compliance dissonance (DissF) regarding attitude-object attributions. The first modification is that we must differentiate the elements of the set *Cognition* into three subsets, namely, *thought, behavior*, and *emotion*, because van Overwalle and Jordens follow the view of three components on attitudes from Rosenberg and Hovland (1960): cognitions, behaviors, and emotions. Instead of using "cognition," this dissertation uses "thought" in order to avoid confusion with its super-set *Cognition*. The set *thought* contains (cognitions of) objects (such as toy, punishment, etc.), the set *behavior* contains (cognitions of) behaviors (such as play, read, etc.) and the set *emotion* contains (cognition of) emotions (such as happy, sad, etc.). The relation between these three new subsets and the set *Cognition* in the $\mathbf{M_p}$(DissF) can be described as follows:

thought $\subseteq$ Cognition

behavior $\subseteq$ Cognition

emotion $\subseteq$ Cognition

Forcecom $\subseteq$ thought

The second modification of $\mathbf{M_p}$(DissF) is defining the set *attitude*, which characterizes the relation between cognition of object and emotion or between cognition of object and behavior.

attitude $\subseteq$ (thought' behavior) $\cup$ (thoughtə emotion)

The set of attitudes contains dissonant and consonant relations between cognitions.

attitude $\subseteq$ Disscog $\cup$ Conscog

In the theory-element of the forced compliance dissonance **T**(DissF) the set *Forcecom*, as a subset of *thought*, also plays a role in defining the set *subattitude* as a subset of *attitude*.

subattitude $\subseteq$ (Forcecom $\times$ behavior) $\cup$ (Forcecom $\times$ emotion)

Through these modifications, the $\mathbf{M_p}$(DissF for Connectionist) can be characterized as follows:

DVII-9: x is a potential model of the forced compliance dissonance ($x \in \mathbf{M_p}$(DissF for Connectionist)) iff there exist Time, Rawcog, Cognition, thought, behavior, emotion, attitude, attint, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, subattitude, subattint, attidiff, imp, reward such that:

(1) $x = \langle$Time, Rawcog, Cognition, thought, behavior, emotion, attitude, attint, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, subattitude, subattint, attidiff, imp, reward$\rangle \in \mathbf{M_p}$(DissF for Connectionist)

(Let x be a potential model of the theory of forced compliance dissonance for the connectionist model)

(2) Time is a finite, non-empty set of points of time

(3) Rawcog is a finite, non-empty set of raw elements of cognition

(4) Cognition $\subseteq$ Rawcog$\times$ time (actual elements of cognition)

(5) thought $\subseteq$ Cognition (cognitions of an object)

(6) behavior $\subseteq$ Cognition (cognitions of behavior)

(7) emotion $\subseteq$ Cognition (cognitions of emotion)

(8) attitude $\subseteq$ (thought $\times$ behavior) $\cup$ (thought $\times$ emotion)

(attitudes to object consist of the relation between object and behavior or emotion)

(9) attint: attitude $\rightarrow IR_0$ (the intensity of attitude)

(10) Disscog $\subseteq$ Cognition $\times$ Cognition

(dissonant cognitions such that Disscog $\subset$ attitude)

(11) Conscog $\subseteq$ Cognition $\times$ Cognition

(consonant cognitions such that Conscog$\subset$attitude)

(12) Disscog $\cap$ Conscog = $\varnothing$

(13) pairdiss := Disscog $\rightarrow$ IR_0^+

(dissonance within pairs)

(14) paircons := Conscog $\rightarrow$ IR_0^+

(consonance within pairs)

(15) pairimp := (Disscog $\cup$ Conscog) $\rightarrow$ IR_0^+

(importance of pairs)

(16) diss := Cognition $\rightarrow$ IR_0^+

(magnitude of dissonance)

(17) redpress := Cognition $\rightarrow$ IR_0^+

(dissonance reduction pressure)

(18) confl(c_{it}) := $\Sigma_{(cit,\ ckt)\in Disscog}$ pairimp(c_{it}, c_{kt})

(degree of conflict)

(19) suppo(c_{it}) := $\Sigma_{(cit,\ ckt)\in Conscog}$ pairimp(c_{it}, c_{kt})

(degree of support)

(20) Forcecom $\subseteq$ thought

(cognitions on counter-attitudinal behavior)

(21) subattitude $\subseteq$ (Forcecom $\times$ behavior) $\cup$ (Forcecom $\times$ emotion) (attitude to cognition on counter-attitudinal behavior)

(22) subattint := subattitude $\rightarrow$ IR_0

(intensity of attitude to cognition on counter-attitudinal behavior)

(23) attidiff := Forcecom $\rightarrow$ IR (attitudinal difference)

(24) imp := Cognition $\rightarrow$ IR_0^+ (importance of cognition)

(25) reward := Forcecom $\rightarrow$ IR_0^+

(magnitude of reward or punishment)

With this modification of the potential models of the forced compliance dissonance ($\mathbf{M_p}$(DissF)), we also should modify the actual models of the forced compliance dissonance **M**(DissF).

DVII-10: x is an actual model of the forced compliance dissonance (x $\in$ **M**(DissF for Connectionist)) iff there exist Time, Rawcog, Cognition, thought, behavior, emotion, attitude, attint, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, subattitude, subattint, attidiff, imp, reward such that:

(1) x = ⟨Time, Rawcog, Cognition, thought, behavior, emotion, attitude, attint, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, subattitude, subattint, attidiff, imp, reward⟩ $\in$ $\mathbf{M_p}$(DissF for Connectionist) (Let x be a potential model of the theory of forced compliance dissonance for the connectionist model)

(2) attitude = $\{(c_{it},c_{jt})|\ (c_{it}\in$thought $\wedge$ $c_{jt}\in$behavior) $\vee$ ($c_{it}\in$thought $\wedge$ $c_{jt}\in$emotion)$\}$

(attitude objects are pairs between the thought of objects and the behavior to objects or pairs between the thought of objects and the emotion to objects)

(3) Disscog $\cup$ Conscog $\subseteq$ attitude

(According to the connectionist model: the cognitive dissonance or cognitive consonance are about attitude objects)

(4) For all (c_{it},c_{jt}), (c_{ku}, c_{lu}) $\in$Disscog: if pairimp (c_{it},c_{jt}) < pairimp(c_{ku},c_{lu}), then pairdiss(c_{it},c_{jt})<pairdiss(c_{ku}, c_{lu})

(If the importance of the pair c_{ku} and c_{lu} is greater than the importance of the pair c_{it} and c_{jt}, then the dissonance of the pair c_{ku} and c_{lu} is greater than the dissonance of the pair c_{it} and c_{jt})

(5) For all (c_{it}, c_{jt}), (c_{ku}, c_{lu}) $\in$Conscog: if pairimp (c_{it},c_{jt}) < pairimp(c_{ku},c_{lu}), then paircons(c_{it},c_{jt})<paircons(c_{ku}, c_{lu})

(If the importance of the pair c_{ku} and c_{lu} is greater than the importance of the pair c_{it} and c_{jt}, then the consonance of the pair c_{ku} and c_{lu} is greater than the consonance of the pair c_{it} and c_{jt})

(6) For all c_{it}, c_{ju} ∈Cognition: if confl(c_{it})/(confl(c_{it}) + suppo (c_{it})) < confl(c_{ju})/(confl(c_{ju}) + suppo (c_{ju})), then diss(c_{it}) < diss(c_{ju}).

(If the proportion between the degree of conflict of c_{ju} and the sum of the importance of c_{ju} is greater than the proportion between the degree of conflict of c_{it} and the sum of the importance of c_{it}, then the dissonance of c_{ju} is greater than the dissonance of c_{it})

(7) For all c_{it}, c_{ju} ∈ Cognition: If diss(c_{it}) < diss(c_{ju}), then redpress(c_{it})<redpress(c_{ju}).

(If the dissonance of c_{ju} is greater than the dissonance of c_{it}, then the attempt to reduce c_{ju} will be greater than the attempt to reduce c_{it})

(8) Subattitude := {(c_{it}, c_{jt})| (c_{it} ∈ Forcecom ∧ c_{jt} ∈ behavior) ∨ (c_{it} ∈ Forcecom ∧ c_{jt} ∈ emotion)}

(According to the connectionist model: the *subattitude* objects are pairs of *Forcecom* and *behavior* towards the object or pairs of *Forcecom* and *emotion* towards the object)

(9) For all c_{it} , c_{ju} ∈ Forcecom:

if$_{cp}$ imp(c_{it})<imp(c_{ju}) or reward(c_{it})>reward(c_{ju}),

then$_{p}$ diss(c_{it})< diss(c_{ju}).

(the more important the opinions or the behavior involved and the smaller the promised reward or threatened punishment, the greater is the magnitude of dissonance created)

(10) For all c_{it}, c_{ju}, c_{it+}, c_{ju+} ∈ Forcecom with t<t+, u<u+:

if$_{cp}$ 0 < redpress(c_{it}) < redpress(c_{ju})

then$_{p}$ 0 > imp(c_{it+}) - imp(c_{it}) > imp(c_{ju+}) - imp(c_{ju})

or 0 < reward(c_{it+}) - reward(c_{it}) < reward(c_{ju+}) - reward(c_{ju})

or 0 > attidiff(c_{it+}) - attidiff(c_{it}) > attidiff(c_{ju+}) - attidiff(c_{ju}).

(Pressure to reduce dissonance may be manifested in reducing the importance or value of the behavior and opinion involved, enhancing the subjective magnitude of the promised reward or threatened punishment, and a change of private opinion following public behavior)

The modification of the potential models of the forced compliance dissonance ($\mathbf{M_p}$(DissF)) leads to a modification of the partial potential models of the forced compliance dissonance ($\mathbf{M_{pp}}$(DissF)). The partial potential model of the forced compliance dissonance can be defined by omitting *attint, pairdiss, paircons, diss, redpress, subattint* because they are the **T**-theoretical elements (see also Chapter 3) as follows:

DVII-11: y is a partial potential model of the forced compliance dissonance for connectionist (y $\in$ $\mathbf{M_{pp}}$(DissF for Connectionist)) iff there exists x such that:

(1) x = ⟨Time, Rawcog, Cognition, thought, behavior, emotion, attitude, attint, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, subattitude, subattint, attidiff, imp, reward⟩$\in$$\mathbf{M_p}$(DissF for Connectionist)

(2) attint, pairdiss, paircons, diss, redpress, subattint are **T**-theoretical.

(3) y = ⟨Time, Rawcog, Cognition, thought, behavior, emotion, attitude, Disscog, Conscog, pairimp, Forcecom, subattitude, attidiff, imp, reward⟩$\in$$\mathbf{M_{pp}}$(DissF for connectionist)

This **T**(DissF for Connectionist) can be seen as a kind of specialization of both original theories – **T**(DissB) and **T**(DissF) – because of the additional requirements. Now we can continue with the modeling and analysis of intertheoretical relations between the theory-element of the forced compliance dissonance (**T**(DissF for Connectionist)) and the unified

theory-element of the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule (**T**(RP+2L-FFNN+DR for Connectionist)) on the connectionist model.

7.3.3. Modeling the Intertheoretical Connections between both Theory-Elements for the Connectionist Model

The connectionist simulation uses only four neurons, which are divided into two layers. They represent the cognitions considered at the time of the simulation. The input neurons represent the thought of a toy and the thought of a threat. The presence and importance of a toy are represented by an input: no toy = 0 and the presence of toy = 1. The presence and the degree of a threat are represented by an input: no threat = 0, mild threat = 0.5, severe threat =1. The output neurons represent behavior and emotion. The first output neuron represents the existence and the kind of behavior: 0 for not play, 1 for play. The second output neuron represents the emotion: -1 for unhappy, 0 for no emotion, +1 for happy. The connections between input neuron and output neuron represent attitude to the toy (*attitude*) or attitude to threat (*subattitude*). The connection weight represents the intensity of attitude to the toy (*attint*) or the intensity of attitude to threat (*subattint*). The dissonance is represented by *Error* in this simulation, which will be reduced by applying the delta rule.

However, in building the structuralist model of intertheoretical relations between the forced compliance dissonance (DissF) and the neural network in the connectionist model, more 'general' models of such intertheoretical connections will be built because this specific case can be expanded for a larger number of cognitions. Let us begin with the intertheoretical connections between the potential models of the forced compliance dissonance for the consonance model ($\mathbf{M_p}$(DissF for connectionist)) and the unifying potential models of the Rosenblatt perceptron, the two layers feed-forward neural network and the delta rule

for the consonance model (**M**$_{\mathbf{p}}$(RP + 2L-FFNN + DR for Connectionist)), which can be defined as follows:

DVII-12: Λ is a collection of determining links between **M**$_{\mathbf{p}}$(DissF for connectionist) and **M**$_{\mathbf{p}}$(RP + 2L-FFNN + DR for Connectionist), e_1 is an echelon set of **M**$_{\mathbf{p}}$(DissF for connectionist) and e_2 is an echelon set of **M**$_{\mathbf{p}}$(RP + 2L-FFNN + DR for Connectionist), and E is a set of entailment links connecting both echelons sets iff there exists x_1, x_2 , λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, λ_7, λ_8, λ_9, λ_{10}, λ_{11}, λ_{12}, λ_{13}, λ_{14} such that:

(1) x_1= ⟨Time, Rawcog, Cognition, thought, behavior, emotion, attitude, attint, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress, Forcecom, subattitude, subattint, attidiff, imp, reward⟩ ∈ **M**$_{\mathbf{p}}$(DissF for Connectionist)

(Let x_1 be a potential model of the forced compliance dissonance (**M**$_{\mathbf{p}}$(DissF)))

(2) x_2= ⟨ N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, η, fnet, fact, fout, net_n, act_n, out_n, Error ⟩ ∈ **M**$_{\mathbf{p}}$(RP + 2L-FFNN + DR for Connectionist)

(Let x_2 be a potential model of the unified model of the two-layers feed-forward neural network, the Rosenblatt's Perceptron, and the delta rule for the connectionist model (**M**$_{\mathbf{p}}$(RP + 2L-FFNN + DR for Connectionist)))

(3) $\lambda_1 \subseteq$ Cognition × N

(λ_1 is a determining link – x is (seen to be) identical with y – that connects the set *Cognition* from **M**$_{\mathbf{p}}$(DissF) to the set of neurons (*N*) from **M**$_{\mathbf{p}}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(4) $\lambda_2 \subseteq$ thought × N_{in}

(λ_2 is a determining link – x is (seen to be) identical with y – that connects the set *thought* from **M**$_{\mathbf{p}}$(DissF) to the set of Input Neurons

(N_{in}) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(5) $\lambda_3 \subseteq \text{Forcecom} \times N_{in}$

(λ_3 is a determining link – x is (seen to be) identical with y – that connects the set *Forcecom* from $\mathbf{M_p}$(DissF) to the set of Input Neurons (N_{in}) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(6) $\lambda_4 \subseteq \text{behavior} \times N_{out}$

(λ_4 is a determining link – x is (seen to be) identical with y – that connects the set *behavior* from $\mathbf{M_p}$(DissF) to the set of Output Neurons (N_{out}) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(7) $\lambda_5 \subseteq \text{emotion} \times N_{out}$

(λ_5 is a determining link – x is (seen to be) identical with y – that connects the set *emotion* from $\mathbf{M_p}$(DissF) to the set of Output Neurons (N_{out}) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(8) $\lambda_6 \subseteq \text{attitude} \times C$

(λ_6 is a determining link – x is (seen to be) identical with y – that connects the set *attitude* from $\mathbf{M_p}$(DissF) to the set of connections (C) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(9) $\lambda_7 \subseteq \text{subattitude} \times C$

(λ_7 is a determining link – x is (seen to be) identical with y – that connects the set *subattitude* from $\mathbf{M_p}$(DissF) to the set of connections (C) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(10) $\lambda_8 \subseteq \text{attint} \times W$

(λ_8 is a determining link – x is (seen to be) identical with y – that connects the set *attint* from $\mathbf{M_p}$(DissF) to the set of connection weights (*W*) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(11) $\lambda_9 \subseteq \text{subattint} \times \text{W}$

(λ_9 is a determining link – x is (seen to be) identical with y – that connects the set *subattint* from $\mathbf{M_p}$(DissF) to the set of connection weights (*W*) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(12) $\lambda_{10} \subseteq \text{imp} \times \text{Inp}$

(λ_{10} is a determining link – x is (seen to be) identical with y – that connects the set of important of cognition (*imp*) from $\mathbf{M_p}$(DissF) to the set of inputs (*Inp*) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(13) $\lambda_{11} \subseteq \text{reward} \times \text{Inp}$

(λ_{11} is a determining link – x is (seen to be) identical with y – that connects the set *Reward* from $\mathbf{M_p}$(DissF) to the set of inputs (*Inp*) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(14) $\lambda_{12} \subseteq \text{behavior} \times \text{Out}$

(λ_{12} is a determining link – x is (seen to be) identical with y – that connects the set *behavior* from $\mathbf{M_p}$(DissF) to the set of desired outputs (*Out*) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(15) $\lambda_{13} \subseteq \text{emotion} \times \text{Out}$

(λ_{13} is a determining link – x is (seen to be) identical with y – that connects the set *emotion* from $\mathbf{M_p}$(DissF) to the set of desired outputs (*Out*) from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(16) $\lambda_{14} \subseteq \text{diss} \times \text{Error}$

(λ_{14} is a determining link – x is (seen to be) identical with y – that connects the set of dissonance (*diss*) from $\mathbf{M_p}$(DissF) to the set *Error* from $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) and a bijective relation)

(17) $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8, \lambda_9, \lambda_{10}, \lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{14}\}$

(Then Λ is a collection of determining links between **T**(DissF for connectionist) and **T**(RP + 2L-FFNN + DR for Connectionist))

(18) $e_1 = \langle$ Cognition, thought, behavior, emotion, attitude, attint, Forcecom, subattitude, subattint, diss, imp, reward$\rangle \in$ echelon set of $\mathbf{M_p}$(DissF for Connectionist)

(e_1 is an echelon set of $\mathbf{M_p}$(DissF for connectionist))

(19) $e_2 = \langle$ N, N_{in}, N_{out}, C, W, Inp, Out, Error $\rangle \in$ echelon set of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist)

(e_2 is an echelon set of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(20) $E = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8, \lambda_9, \lambda_{10}, \lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{14}\}$

(E is a set of entailment links between e_1 and e_2)

The interpreting links in the level of the partial potential models are characterized in order to determine the local empirical claims and the intended applications. Based on D VIII-7 and VIII-8 in BMS, pp. 398–400 the interpreting links for the reduction of DissF for connectionist by RP + 2L-FFNN + DR for Connectionist can be determined as follows:

DVII-13: E^*(DissF) = $\{l_1, l_2, l_3, l_4, l_5\ l_6, l_7, l_8, l_9, l_{10}, l_{11}, l_{12}, l_{13}, l_{14}\}$ is a collection of interpreting links, where **T**(DissF for Connectionist) is interpreted by **T**(RP + 2L-FFNN + DR for Connectionist) in the connectionist model iff there exist x_1, x_2, e_1, e_2, y_1, f_1 such that:

(1) $x_1 = \langle$Time, Rawcog, Cognition, thought, behavior, emotion, attitude, attint, Disscog, Conscog, pairdiss, paircons, pairimp, diss, redpress,

Forcecom, subattitude, subattint, attidiff, imp, reward⟩ ∈ $\mathbf{M_p}$(DissF for Connectionist)

(x_1 is a potential model of the forced compliance dissonance for the connectionist ($\mathbf{M_p}$(DissF for Connectionist)))

(2) x_2= ⟨N, N_{in}, N_{out}, IN, IR, C, pred(n), succ(n), W, B, W_0, Inp, Out, L, η, fnet, fact, fout, net_n, act_n, out_n, Error⟩ ∈ $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist)

(x_2 is a potential model of the unified Perceptron, 2-layers feed forward neural network, and delta-rule ($\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist)))

(3) e_1 = ⟨Cognition, thought, behavior, emotion, attitude, attint, Forcecom, subattitude, subattint, diss, imp, reward⟩ ∈ echelon set of $\mathbf{M_p}$(DissF for Connectionist)

(e_1 is an echelon set of $\mathbf{M_p}$(DissF for Connectionist))

(4) e_2 = ⟨N, N_{in}, N_{out}, C, W, Inp, Out, Error⟩ ∈ echelon set of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist)

(e_2 is an echelon set of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(5) E = {λ_1, λ_2, λ_3, λ_4, λ_5 λ_6, λ_7, λ_8, λ_9, λ_{10}, λ_{11}, λ_{12}, λ_{13}, λ_{14}}

(E is a set of entailment links between e_1 and e_2)

(6) y_1 = ⟨Time, Rawcog, Cognition, thought, behavior, emotion, attitude, Disscog, Conscog, pairimp, Forcecom, subattitude, attidiff, imp, reward⟩ ∈ $\mathbf{M_{pp}}$(DissF for connectionist)

(y_1 is the partial potential model of the forced compliance dissonance for connectionist ($\mathbf{M_{pp}}$(DissF for Connectionist)))

(7) E*(DissF) = {l_1, l_2, l_3, l_4, l_5 l_6, l_7, l_8, l_9, l_{10}, l_{11}, l_{12}, l_{13}, l_{14}} ∈ the set of interpreting links,where {l_i = ⟨x',y⟩| x'∈$\mathbf{M_p}$(**T'**), y∈$\mathbf{M_{pp}}$*(**T**) and there is x∈$\mathbf{M_p}$(**T**) such that ⟨x',x⟩∈(**T'**, **T**) and r*(x)=y}

(E*(DissF for Connectionist) is the collection of interpreting links)

(8) $l_1 = \{\langle n_j, \text{cognition}_i\rangle | \ N(n_j) \wedge \text{Cognition}(\text{cognition}_i) \rightarrow R(n_j, \text{cognition}_i)\}$ (R(x,y) = x interprets y, and R is bijective)

(l_1 is the interpreting link, that interprets each cognition in DissF for Connectionist being represented by a neuron in RP + 2L-FFNN + DR for Connectionist. In the simulation there are 4 neurons – 2 input neurons and 2 output neurons – represent toy, threat, play and happy. The *Cognition* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *N* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(9) $l_2 = \{\langle n_{in(j)}, \text{thou}_i\rangle | \ N_{in}(n_{in(j)}) \wedge \text{thought}(\text{thou}_i) \rightarrow R(n_{in(j)}, \text{thou}_i)\}$

(l_2 is the interpreting link, that interprets each *thought* in DissF for Connectionist being represented by an input neuron in RP + 2L-FFNN + DR for Connectionist. In the simulation $n_{in(1)}$ = toy. The *thought* here is an element of $\mathbf{M_{pp}}$(DissF), whereas N_{in} is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(10) $l_3 = \{\langle n_{in(j)}, \text{forcecom}_i\rangle | \ N_{in}(n_{in(j)}) \wedge \text{Forcecom}(\text{forcecom}_i) \rightarrow R(n_{in(j)}, \text{forcecom}_i)\}$

(l_3 is the interpreting link, that interprets each *forcecom* in DissF for Connectionist being represented by an input neuron in RP + 2L-FFNN + DR for Connectionist. In the simulation $n_{in(2)}$ = threat. The *Forcecom* here is an element of $\mathbf{M_{pp}}$(DissF), whereas N_{in} is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(11) $l_4 = \{\langle n_{out(j)}, \text{behav}_i\rangle | \ N_{out}(n_{out(j)}) \wedge \text{behavior}(\text{behav}_i) \rightarrow R(n_{out(j)}, \text{behav}_i)\}$

(l_4 is the interpreting link, that interprets each behavior in DissF for Connectionist being represented by output neurons in RP + 2L-FFNN + DR for Connectionist. In the simulation $n_{out(1)}$ = play. The *behavior* here is an element of $\mathbf{M_{pp}}$(DissF), whereas N_{out} is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(12) $l_5 = \{\langle n_{out(j)}, emo_i\rangle | N_{out}\ (n_{out(j)}) \wedge emotion(emo_i) \rightarrow R(n_{out(j)}, emo_i)\}$

(l_5 is the interpreting link, that interprets each emotion in DissF for Connectionist being represented by output neurons in RP + 2L-FFNN + DR for Connectionist. In the simulation $n_{out(2)}$ = happy. The *emotion* here is an element of $\mathbf{M_{pp}}$(DissF), whereas N_{out} is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(13) $l_6 = \{\langle c_j, att_i\rangle | C(c_j) \wedge attitude(att_i) \rightarrow R(c_j, att_i)\}$

(l_6 is the interpreting link, that interprets attitude in DissF for Connectionist being represented by an inter-neuronal connection in RP + 2L-FFNN + DR for Connectionist. The *attitude* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *C* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(14) $l_7 = \{\langle c_j, subatt_i\rangle | C(c_j) \wedge subattitude(subatt_i) \rightarrow R(c_j, subatt_i)\}$

(l_7 is the interpreting link, that interprets *subattitude* in DissF for Connectionist being represented by an inter-neuronal connection in RP + 2L-FFNN + DR for Connectionist. The *subattitude* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *C* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(15) $l_8 = \{\langle w_j, atti_i\rangle | W(w_j) \wedge attint(atti_i) \rightarrow R(w_j, atti_i)\}$

(l_8 is the interpreting link, that interprets *attint* in DissF for Connectionist being represented by connection weights in RP + 2L-FFNN + DR for Connectionist. The *attint* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *W* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(16) $l_9 = \{\langle w_j, subatti_i\rangle | W(w_j) \wedge subattint(subatti_i) \rightarrow R(w_j, subatti_i)\}$

(l_9 is the interpreting link, that interprets *subattint* in DissF for Connectionist being represented by connection weights in RP + 2L-

FFNN + DR for Connectionist. The *subattint* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *W* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(17) $l_{10} = \{\langle inp_j, imp_i\rangle | \; Inp(inp_j) \wedge imp(imp_i) \rightarrow R(inp_j, imp_i)\}$

(l_{10} is the interpreting link, that interprets *imp* in DissF for Connectionist being represented by the input in RP + 2L-FFNN + DR for Connectionist. In the simulation the possibilities are {(not present, 0), (present, 1)}. The *imp* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *Inp* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(18) $l_{11} = \{\langle inp_j, rew_i\rangle | \; Inp(inp_j) \wedge reward(rew_i) \rightarrow R(inp_j, rew_i)\}$

(l_{11} is the interpreting link, that interprets reward in DissF for Connectionist being represented by the input in RP + 2L-FFNN + DR for Connectionist. In the simulation the possibilities are {(not present, 0), (mild, 0.5), (severe, 1)}. The *Reward* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *Inp* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(19) $l_{12} = \{\langle out_j, behav_i\rangle | \; Out\ (out_j) \wedge behavior(behav_i) \rightarrow R(out_j, behav_i)\}$

(l_{12} is the interpreting link, that interprets behavior in DissF for Connectionist being represented by the output in RP + 2L-FFNN + DR for Connectionist. In the simulation, the possibilities are {(not play, 0), (play, 1)}. The *behavior* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *Out* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(20) $l_{13} = \{\langle out_j, emo_i\rangle | \; Out(out_j) \wedge emotion(emo_i) \rightarrow R(out_j, emo_i)\}$

(l_{13} is the interpreting link, that interprets emotion in DissF for Connectionist being represented by the output in RP + 2L-FFNN +

DR for Connectionist. In the simulation the possibilities are {(unhappy, -1), (neutral, 0), (happy, 1)}. The *emotion* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *Out* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(21) $l_{14} = \{\langle error_i, diss_i\rangle | \text{ Error}(error_j) \wedge \text{diss}(diss_i) \rightarrow R\ (error_j, diss_i)\}$

(l_{14} is the interpreting link, that interprets dissonance in DissF for Connectionist represented by the Error in RP + 2L-FFNN + DR for Connectionist. The *dissonance* here is an element of $\mathbf{M_{pp}}$(DissF), whereas *Error* is an element of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist))

(22) $f_1 = \langle$Cognition, thought, behavior, emotion, attitude, Forcecom, subattitude, imp, reward$\rangle \in$ an echelon subset of $\mathbf{M_{pp}}$(DissF) by applying function r*: $e_1 \rightarrow f_1$, that mapping E*(DIssF) from $\mathbf{M_p}$ (DissF) to $\mathbf{M_{pp}}$(DissF)

(f_1 is an echelon subset of $\mathbf{M_{pp}}$(DissF) with respect to E*(DissF))

Based on D VIII-9 and DVIII-10 in BMS the f_1 can be determined as the set of empirical claims of the intertheoretical reduction of the theory of forced compliance dissonance by RP + 2L-FFNN + DR for Connectionist. The empirical claims of this intertheoretical reduction refer to the sets *cognition, thought, behavior, emotion, attitude, attint, Forcecom, subattitude, imp, Reward* in the partial potential models of the forced compliance dissonance for connectionist. They are the specialization of the original forced compliance dissonance by adding new categories in the concept of cognition and the relation between them. The interpreting links E*(DissF for Connectionist) interpret these concepts being represented by some concepts from RP + 2L-FFNN + DR for Connectionist.

A Summary of this Chapter. A structuralist model of the intertheoretical connections between the adapted forced compliance dissonance (DissF for connectionist) and the adapted two-layers feed-forward perceptrons with the delta rule for the connectionist model (RP + 2L-FFNN + DR for Connectionist) was built. In this intertheoretical reduction, only the echelon partial subset of $\mathbf{M_p}$(DissF for connectionist) is connected to the echelon partial subset of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist). Our model can be presented in Figure 7.3.

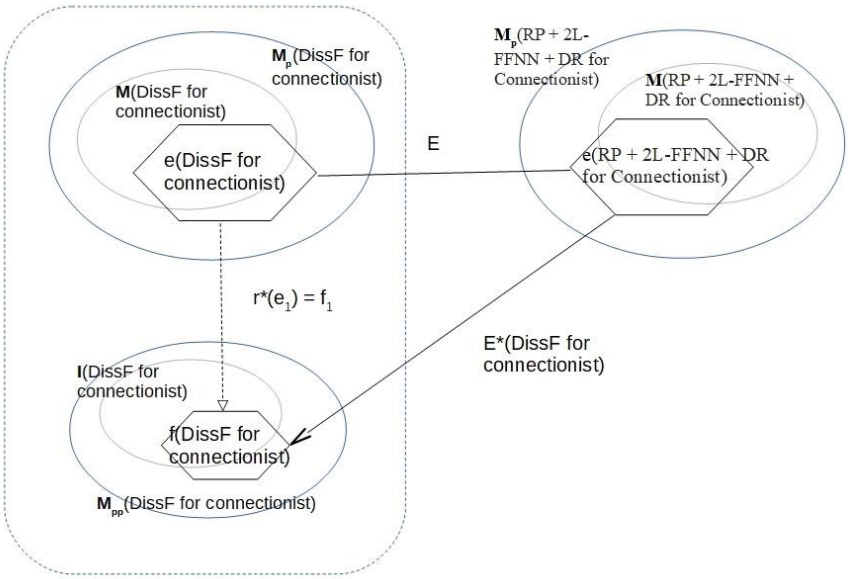


Figure 7.3. A structuralist modeling of intertheoretical reduction between the adapted forced compliance dissonance (DissF for connectionist) and the adapted two-layers feed-forward neural network for the connectionist model (RP + 2L-FFNN + DR for Connectionist). e(DissF for connectionist) is the echelon partial subset of $\mathbf{M_p}$(DissF for Connectionist) and e(RP + 2L-FFNN + DR for Connectionist) is the echelon partial subset of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist). Both are connected by a set of entailment Links (E). The interpreting links with local at DissF for connectionist (E*(DissF for Connectionist)) connect the set of empirical claim of this intertheoretical reduction at DissB and the echelon partial subset of $\mathbf{M_p}$(RP + 2L-FFNN + DR for Connectionist) as the reducing theory.

In the first part of this chapter it has been discussed about the so-called V-pattern. Now, having discussed the intertheoretical connections, we want to look back to the notion of V-pattern and how it works in the interdisciplinary context – especially if the relations between theories in the disciplines involved are very complex or contain the V-pattern too. The case of the connectionist model is interesting because both combined theories have a kind of V-pattern. In this model, both V-patterns are connected by a link with one of them serving as the mainboard theory. We are dealing with a multi-level V-pattern, which can be depicted in graph theory as follows:

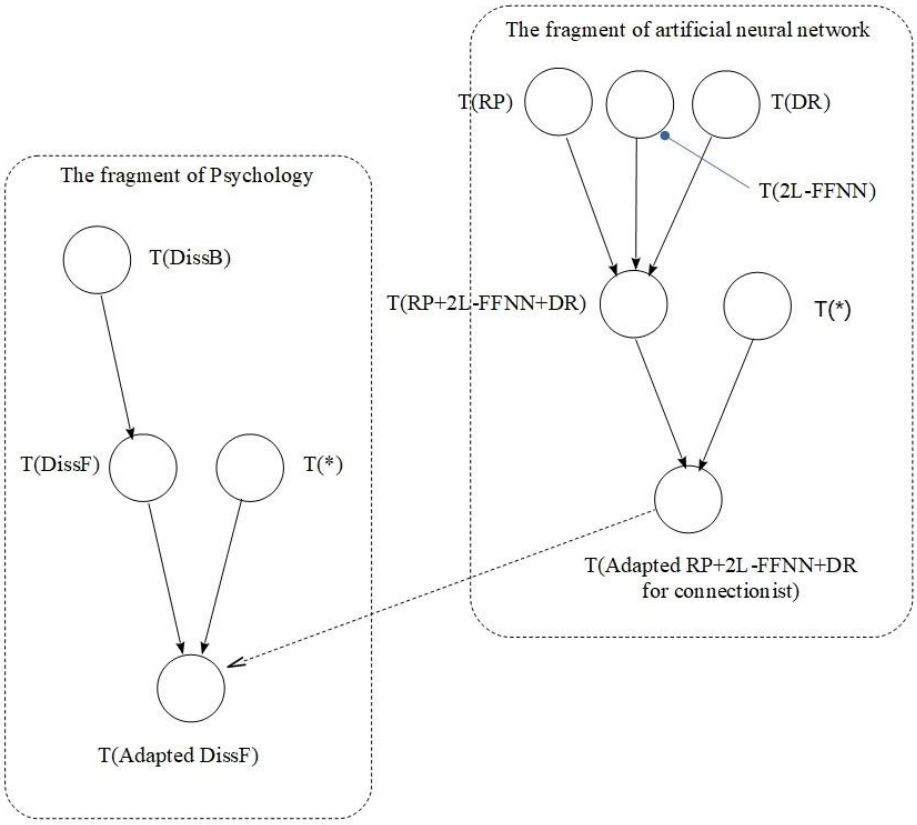


Figure 7.4. The pattern of intertheoretical relations between the adapted theory of forced compliance dissonance and the adapted unifying theory of the perceptron, the two-layers feed-forward neural network and the delta-rule for connectionist

Chapter 8
The Contribution of This Research for Philosophy of Science, Cognitive Science, and Interdisciplinary Practices

This chapter will discuss the significance of this research for the relevant scientific fields, namely the philosophy of science, cognitive science, and interdisciplinary practices. In the philosophy of science, this research contributes with further development of the structuralist theory of science not only as a first implementation of this theory in modeling the intertheoretical connections in interdisciplinary fields, and not only in the adjustment of the definition of specialization provided in Chapter 4 but also in the formulation of a new specialization of the notion of a theory-holon related to the combination of theories in scientific practices that will be the first topic in this chapter. The second topic in this chapter will be a comparison between the results of modeling intertheoretical reduction as having been done in the last three chapters with the generalized Nagel-Schaffner approach. This chapter will also discuss the idea of the unity of science and show how a program for the unity of science is possible. For cognitive science and other interdisciplinary practices, this dissertation is the first attempt at modeling and analyzing intertheoretical connections between theories from various disciplines by implementing formal methods to show how a successful intertheoretical combination works logically.

8.1. A Further Development in the Structuralist Theory of Science

The modeling of intertheoretical connections in the last three chapters shows the usefulness of the structuralist theory of science in mapping the intertheoretical connections between two or more theories within one single discipline or between various disciplines. The structuralist

theory of science cannot only be used to model the intertheoretical relations between the concepts of the theories but also the modifications needed such that the connected theories work as expected (planned). By these models, we can discover a new pattern of connections, a strategy to connect or combine several theories, and a recipe to build a more complex model or theory, and new association links.

8.1.1. The V-Pattern of Intertheoretical Connections and A Strategy to Build A More Complex Model by Unifying Several Theories

Balzer, Moulines, and Sneed have discussed several types of intertheoretical connections and formulated their formal definitions. This dissertation is an attempt to implement those definitions to build and analyze several real cases of intertheoretical connections in interdisciplinary fields, especially in cognitive science. This attempt provides a chance to learn more about the intertheoretical connections – and admittedly, there will be more to learn in the future. In Chapter 5, an intertheoretical relation between the Festinger theory of cognitive dissonance and the computational neuroscientific theory has been modeled and analyzed concerning the interdisciplinary research on the *dorsal Anterior Cingulate Cortex* conducted by van Veen et al. (2009). This modeling and analysis specify in a formal way how the concepts of both theories connect and how to formulate the local intended applications or empirical claims of the connections, which shows how far the connected theories can explain their intended phenomena. However, this modeling is the simplest of our models.

In Chapters 6 and 7 author attempts to model and analyze more detailed and more complicated intertheoretical connections, namely intertheoretical connections between two or more theories in a simulation. A simulation is not as simple as an explanation because a simulation must add several additional requirements to mimic the simulated phenomena. It is a fact that the original form of the theories involved is not (always) ready for

such an intended application. In Chapter 6, the Hopfield model of the recurrent neural network is not ready for cognitive dissonance simulation. It needs some modifications or adjustments to build a simulation according to the consonance model. The result of such change is seen as a specialization of the original theory. The connectionist simulation gives the same or even a more complicated situation in Chapter 7. Not only some adjustments are needed, but also a combination of several theory-elements at once, namely the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule, to build a new theoretical unity.

Before continuing the discussion about the combination of the three theories, it is necessary to explain why the presentation is made complicated by seeing them as three theories. The Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule usually are applied together in data science and machine learning. Some readers may think that they can be seen simply as one model, and it is not necessary to discuss a combination of theories and their intertheoretical connections. However, such an opinion oversimplifies the real conceptual and methodological situation. Such oversimplification makes some interesting and important points stayed unclear. In this work, they are seen as three independent theories, and three theory-elements are modeled for several reasons. First, these theories are indeed three conceptually and methodologically different theories. In machine learning or data science, people often replace one of these theories with another model or theory. A very common example is that people replace the two-layers feed-forward neural network with a multi-layers feed-forward neural network, but still use the perceptron and the delta rule as its learning rule. They place perceptrons in some additional layers between the input layer and the output layer, and they use the delta rule to train the multi-layers perceptrons (MLP). Second, it leads to the importance of research on intertheoretical connections in general. Notably, this topic is relatively rarely discussed in comparison to the topic of reduction or unity

of science, although there are many types of intertheoretical connections that can be explored and used in scientific practice. Third, it points out the importance of modeling and analyzing intertheoretical connections for modeling complex phenomena. Modeling and analyzing intertheoretical connection can be beneficial not only to evaluate our steps in combining and adjusting the theories implemented but also to formulate a general strategy to build a more complex theory or model to explain complex phenomena, especially in our interdisciplinary age.

Inspired by the idea of the local empirical claim of the theory-holon to interpret the model of intertheoretical connections in Chapter 5, a special pattern for combining scientific theories, called "V-pattern," has been discovered. This name is used because of its graphic representation when determining the local empirical claims or intended applications. And this section will characterize the definition of the V-pattern and explain a strategy to implement it. The V-pattern has the following main features: (1) There exists a theory-element $\mathbf{T_0}$ that serves as the mainboard theory of the V-pattern of intertheoretical relations. The mainboard theory is a theory to which all other theories are being connected in this pattern. The mainboard theory represents a basic model of phenomena that we want to explain through our set of theories in a holon. (2) $\mathbf{T_1, ..., T_n}$ are the connected theories that can be connected one by one to the $\mathbf{T_0}$ with an intertheoretical connection. These connected theories enrich the basic model represented by the $\mathbf{T_0}$ to deliver a more holistic explanation or application. These features can be formally defined as follows:

D VIII-1: v is a V-pattern of an intertheoretical relation iff there exist $\mathbf{T_0}$, **H**, and λ such that:

(1) $v = \langle \mathbf{T_0}, \mathbf{H}, \lambda \rangle \in V$

(2) $\mathbf{T_0}$ is the main board theory-element.

(3) **H** is a set of connected theory-elements to $\mathbf{T_0}$, where **H** has at least one element.

(4) Λ is a non-empty set of intertheoretical connections (links).

(5) For $\mathbf{T_0}$ there are $\mathbf{T_1, T_2, \ldots, T_n} \in \mathbf{H}$ such that: $(\mathbf{T_1}\ \lambda_1\ \mathbf{T_0}) \wedge (\mathbf{T_2}\ \lambda_2\ \mathbf{T_0}) \wedge \ldots \wedge (\mathbf{T_n}\ \lambda_n\ \mathbf{T_0})$, where $\lambda_i \in \Lambda$.

Besides the two main features above, other features can also be described as follows: First, the intertheoretical connections on the class of the potential models can be modeled as the dyadic relations between concepts of the mainboard theory and concepts of additional theory(es) by using the determining links that represent a relation "x is identical with y" unless a certain specific relation defined. The type of intertheoretical connection used here is exclusively the determining link because in combining theories, it is about connections between concepts. Second, the intended application of the V-pattern of intertheoretical relations is local on the mainboard theory $\mathbf{T_0}$. Since the V-pattern is a pattern for theory-holon, the intended applications are local. In the V-pattern the local intended applications are placed on the mainboard theory $\mathbf{T_0}$, whose $\mathbf{T_0}$-non-theoretical concepts are being interpreted by other concepts of other connected theories $\mathbf{T_1, T_2, \ldots, T_n}$ through interpreting links. This relation can be described in a directed acyclic graph as follows:

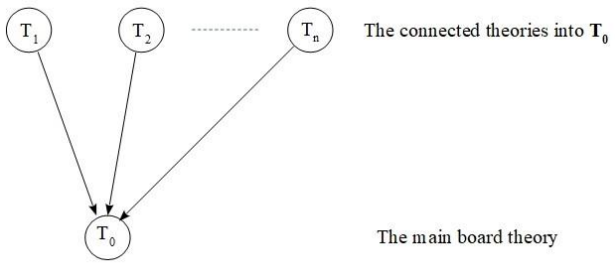


Figure 8.1. The directed acyclic graph of the V-pattern of intertheoretical relation

This V-pattern is a kind of specialization of the general pattern of connected theory-holon as characterized in BMS as D VIII-1. The V-pattern, as a specialization of theory-holon, can be defined as follows:

D VIII-2: **H** is a V-pattern in a theory-holon iff there exist $\mathbf{N_0}$, **N**, and $\mathbf{\Lambda}$, such that $\mathbf{H} = \langle \mathbf{N_0}, \mathbf{N}, \mathbf{\Lambda} \rangle$ and:

(1) $\mathbf{N_0}$ and **N** are non-empty sets of theory-elements. $\mathbf{N_0}$ is the theory-net that contains the mainboard theory ($\mathbf{T_0 \in N_0}$), and **N** is the set of theory-nets that contain theory-elements connected to the mainboard theory.

(2) $\mathbf{\Lambda}$: $\mathbf{N_0 \times N} \rightarrow \cup\{\text{Po}(\mathbf{M_p(T_0)} \times \mathbf{M_p(T)})/\mathbf{T_0 \in N_0}, \mathbf{T \in N}\}$ is a partial function.

(3) For all **T,** there exists $\mathbf{T_0}$: Let $\langle \mathbf{T_0, T} \rangle \in \text{Dom}(\lambda)$, then $\lambda(\mathbf{T_0, T}) \subseteq \mathbf{M_p(T_0) \times M_p(T)}$.

(4) If **N** contains more than one element, then there is $\mathbf{T_0 \in N_0}$, and there is $\mathbf{T \in N}$, such that $\langle \mathbf{T_0, T} \rangle \in \text{Dom}(\lambda)$ or $\langle \mathbf{T, T_0} \rangle \in \text{Dom}(\lambda)$.

(5) For all $\mathbf{T_0 \in N_0, T_1, T_2 \in N}$: Let $\langle \mathbf{T_0, T_1} \rangle \in \text{Dom}(\lambda)$ and $\langle \mathbf{T_0, T_2} \rangle \in \text{Dom}(\lambda)$, then $\langle \mathbf{T_1, T_2} \rangle \in \text{Dom}(\lambda)$.

The V-pattern serves as a tool for helping to build a (more) complex combined theory for explaining and modeling a complex phenomenon or for creating a complex implementation or application. This tool is formulated because scientific theories typically have two features that make it difficult to make a comprehensive application. One of the features of scientific theories is that a theory must be general enough – a scientific theory does not model a specific or single phenomenon, but general phenomena. For example, the Festinger theory of cognitive dissonance explains not only dissonance reduction of one person but also similar processes of many (if not all) persons. Because of this kind of generalization, a scientific theory explains certain phenomena only as far as some aspects are concerned.

Because of these two features, building a comprehensive explanation, model, application, or implementation of a theory of a particularly complex phenomenon needs some modification and combination of some given scientific theories. Chapters 6 and 7 show that building a simulation requires some adjustments to the theories from both disciplines.

For such purposes, the V-pattern can be implemented through the following strategy: (1) We choose a theory about complex phenomena as a mainboard theory. Some considerations for choosing the mainboard theory are as follows: (a) Choose a theory from a discipline explaining the phenomena on which the research focuses. In Chapter 7, the connectionist model focus on the simulation of dissonance reduction. Therefore, the mainboard theory for this combination is the Festinger theory. (b) It would be better if the chosen theory contains most of the concepts we need. The local intended application will be focused on the mainboard theory. (2) We connect other theory-elements to the mainboard theory-element. Through both steps, we will get a schema similar to Figure 8.1. (3) We unite the theories into one single new theory by the following procedure. (a) By combining all the potential models of all theories to be the single unified potential models. (b) By reducing the elements of the unifying potential models by omitting the superfluous or redundant concepts and by transferring all the relations and functions of the omitted elements to the rest identical concepts. (4) There is also the possibility of combining several V-patterns to building a more complex model by following the steps one through three and placing the mainboard theory containing fewer concepts in level one and connecting it to the mainboard theory-element containing more concepts. The pattern can be described as follows:

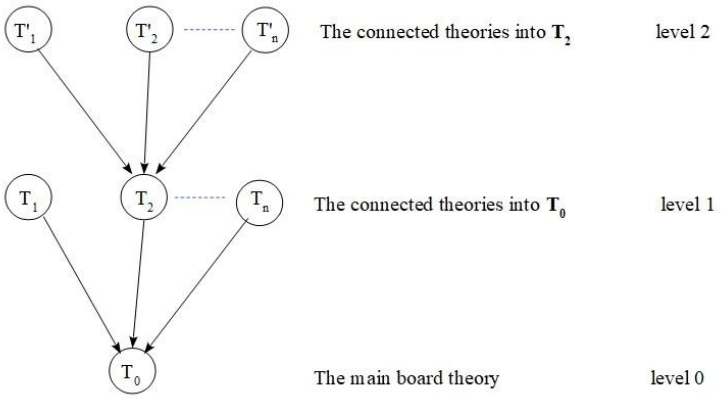


Figure 8.2. The Directed Acyclic Graph of the Combined V-Patterns of Intertheoretical Relation; T_2 Is the Other Main Board Theory-Element, to Which the $\mathbf{T'_1, T'_2, \ldots, T'_n}$ Are Connected. $\mathbf{T_0}$ Is More General than $\mathbf{T_2}$. Therefore, We Have to Put $\mathbf{T_0}$ in the Level 0 and $\mathbf{T_2}$ in the Level 1 Connected to some Term(s) of T_0.

8.1.2. The Unifying Theory-Element

In Chapter 7, a new kind of theory-element is introduced, a combination of several theory-elements into one single theory-element. This model simplifies the model of intertheoretical connections and unifies several connected theories or models to build a more complex theory or model. This new kind of theory-element will be called as unifying theory-element.

A unifying theory-element is essentially a unification of several combined theories, where some or all concepts are connected through a bijective relation "x is identical with y." For unifying those theories, the dyadic intertheoretical connections are sufficient by implementing of the V-pattern and following the strategy laid out in section 8.1.1 above. The unifying theory-element can be characterized as follows:

D VIII-3 : $\mathbf{T^U} = \langle \mathbf{M_p^U}, \mathbf{M^U}, \mathbf{M_{pp}^U}, \mathbf{\Lambda^U}, \mathbf{E^{*u}} \rangle$ is the unifying theory-element between $\mathbf{T^0}$, $\mathbf{T^1}$, …, $\mathbf{T^n}$ iff there exist $\mathbf{T^0} = \langle \mathbf{M_p^0}, \mathbf{M^0}, \mathbf{M_{pp}^0}, \mathbf{E^{*0}} \rangle$, $\mathbf{T^1} = \langle \mathbf{M_p^1}, \mathbf{M^1}, \mathbf{M_{pp}^1} \rangle$, …, $\mathbf{T^n} = \langle \mathbf{M_p^n}, \mathbf{M^n}, \mathbf{M_{pp}^n} \rangle$, $\boldsymbol{\lambda^1}, \boldsymbol{\lambda^2}, \ldots, \boldsymbol{\lambda^n}$, $l^1, l^2, \ldots, l^n$ such that:

(1) $\mathbf{M_p^U} := \mathbf{M_p^0} \cup \mathbf{M_p^1} \cup \ldots \cup \mathbf{M_p^n}$

($\mathbf{M_p^U}$ is the set of potential models of the new unifying theory-element ($\mathbf{T^U}$))

(2) $\mathbf{M^U} := \mathbf{M^0} \cup \mathbf{M^1} \cup \ldots \cup \mathbf{M^n}$

($\mathbf{M^U}$ is the set of actual models of the new unifying theory-element ($\mathbf{T^U}$))

(3) $\mathbf{M_{pp}^U} := r°(\mathbf{M_p^U})$, where r° is a function that project $\mathbf{M_p^U}$ to $\mathbf{M_{pp}^U}$

($\mathbf{M_{pp}^U}$ is the partial potential models of the new unifying theory-element ($\mathbf{T^U}$))

(4) $\mathbf{\Lambda^U} := \boldsymbol{\lambda}^1 \cup \boldsymbol{\lambda}^2 \cup \ldots \cup \boldsymbol{\lambda}^n$, where:

1) $\boldsymbol{\lambda}^1 \subseteq \mathbf{M_p^0}, \mathbf{M_p^1}$

2) $\boldsymbol{\lambda}^2 \subseteq \mathbf{M_p^0}, \mathbf{M_p^2}$

…) ...

n) $\boldsymbol{\lambda}^n \subseteq \mathbf{M_p^0} \times \mathbf{M_p^n}$

($\mathbf{\Lambda^U}$ is the set of the unifying intertheoretical connections (links) that connect $\mathbf{M_p^0}$ to $\mathbf{M_p^i}$, and i = 1, 2, ... n, where:

1) $\boldsymbol{\lambda}^1$ is the set of unifying intertheoretical connections (links) that connect $\mathbf{M_p^0}$ to $\mathbf{M_p^1}$

2) $\boldsymbol{\lambda}^2$ is the set of unifying intertheoretical connections (links) that connect $\mathbf{M_p^0}$ to $\mathbf{M_p^2}$

…

n) $\boldsymbol{\lambda}^n$ is the set of unifying intertheoretical connections (links) that connect $\mathbf{M_p^0}$ to $\mathbf{M_p^n}$)

(5) $\mathbf{E^{*U}} = \mathbf{E^{*0}} := \mathbf{l^1} \cup \mathbf{l^2} \cup \ldots \cup \mathbf{l^n}$, where:

1) $l^1 = \{\langle x^1, y^0\rangle \mid x^1 \in M_p^1 \wedge y^0 \in M_{pp}^0 \wedge x^0 \in M_p^0 \rightarrow \langle x^1, x^0\rangle \in (\mathbf{T^1}, \mathbf{T^0})$ and $r^*(x^0)=y^0\}$

2) $l^2 = \{\langle x^2, y^0\rangle \mid x^2 \in M_p^2 \wedge y^0 \in M_{pp}^0 \wedge x^0 \in M_p^0 \rightarrow \langle x^2, x^0\rangle \in (\mathbf{T^2}, \mathbf{T^0})$ and $r^*(x^0)=y^0\}$

…) …

n) $l^n = \{\langle x^n, y^0\rangle \mid x^n \in M_p{}^n \wedge y^0 \in M_{pp}{}^0 \wedge x^0 \in M_p{}^0 \rightarrow \langle x^n, x^0\rangle \in (\mathbf{T^n}, \mathbf{T^0})$ and $r^*(x^0)=y^0\}$

($\mathbf{E^{*U}}$ is the set of interpreting links that connect $\mathbf{M_{pp}{}^0}$ to $\mathbf{M_p{}^i}$, such that $\mathbf{M_p{}^i}$ is interpreting $\mathbf{M_{pp}{}^0}$, and i = 1, 2, ... n, where:

1) l^1 is the set of interpreting links that connect $\mathbf{M_p{}^1}$ to $\mathbf{M_{pp}{}^0}$, where $\mathbf{T^1}$ interprets $\mathbf{T^0}$.

2) l^2 is the set of interpreting links that connect $\mathbf{M_p{}^2}$ to $\mathbf{M_{pp}{}^0}$, where $\mathbf{T^2}$ interprets $\mathbf{T^0}$.

…

n) l^n is the set of interpreting links that connect $\mathbf{M_p{}^n}$ to $\mathbf{M_{pp}{}^0}$, where $\mathbf{T^n}$ interprets $\mathbf{T^0}$)

The intuitive idea behind this definition is as follows: Suppose there are several theory-elements to combine. The unifying intertheoretical connections implemented here between them are dyadic relations that connect the mainboard theory element ($\mathbf{T^0}$) and the other theory-elements ($\mathbf{T^1, \ldots, T^n}$). Because the theory-elements we want to combine are different, the potential models of each theory-element are also different. Because of this fact, the unifying intertheoretical connections between two theory-elements are a set of determining links. Like the other kinds of intertheoretical connections in the structuralist theory of science, the unifying intertheoretical connections are also the relations between the potential models of both combined theories. Based on these considerations, the $\mathbf{M_p{}^U}$ can be built by unifying the $\mathbf{M_p}$ of the combined theory-elements. This new unifying theory-element will also contain all law-statements of the combined theory-elements. Therefore, the actual models for the unifying theory-element ($\mathbf{M^U(T^U)}$) are defined as the unification of the actual models ($\mathbf{M}$) of the connected theory-elements. The set of unifying intertheoretical connections ($\mathbf{\Lambda}$) connects the elements of all $\mathbf{M_p}$ of the connected theory-

elements. The partial potential models of the unifying theory-element ($\mathbf{M_{pp}^U(T^U)}$) can also be characterized by implementing the function r° that maps the $\mathbf{M_p^U}$ to $\mathbf{M_{pp}^U}$ by omitting the T-theoretical elements of $\mathbf{M_p^U}$.

8.2. Philosophy of Science in General

There are two major issues in the philosophy of science in general to which our modeling has relevance. The first one is the issue of intertheoretical reduction, which dominates the discussion about the relation between theories in the philosophy of science today. Another issue in the philosophy of science that is relevant to the results of this work is the issue of the unity of science.

8.2.1. Intertheoretical Reduction

One of the most influential theories of intertheoretical reduction is the generalized Nagel-Schaffner (GNF) theory of intertheoretical reduction. This theory is an improvement on the original version of Nagel's theory of reduction. The basic idea of Nagel's original account of reduction is relatively simple. A theory T_P reduces to another theory T_F iff the laws of T_P can be deduced from the laws of T_F and some auxiliary assumptions. The auxiliary assumptions are typically idealizations and boundary conditions. Nagel considers two formal conditions for reduction concerning the formal nature of theories. The two conditions for successful reduction are connectability and derivability. The condition of connectability requires that for every theoretical term in T_P there is a theoretical term in T_F corresponding to it. The condition of derivability says that if connectability is satisfied, the laws of T_P can be derived from the laws of T_F plus auxiliary assumptions. For Nagel, there are also two kinds of reduction, namely homogeneous reduction and heterogeneous reduction. In the homogeneous reduction, both theories share the same relevant predicates. Therefore, the

connectability requirement is trivially satisfied. Whereas in heterogeneous reduction, the relevant terms of both theories are not the same.

The case of intertheoretical reduction in interdisciplinary fields is the heterogeneous reduction and is rarely a homogeneous one. Firstly, it is obvious in our cases because we model the intertheoretical reduction, which connects some psychological concepts to some biological entities or mathematical entities. The theories do not share the same relevant terms. Secondly, although our theories sometimes use the same words, they have different basic concepts or relations defined according to their disciplines. For example, let's examine the word "neuron" being used in both neuroscience and artificial neural networks. The meaning of neurons in both fields is very different. In neuroscience, the term 'neuron' refers to cells in the brain and the nervous system of living creatures, whereas the term 'neuron' in artificial neural network refers to an abstract mathematical model of a neuron written in a computer program.

Because of the heterogeneous reduction, the relevant terms of connected theories are not the same. It is impossible to derive the laws of T_P from T_F directly. For making the reduction possible, bridge laws are needed to connect the vocabulary of T_P and T_F by providing 'rules of translation.' The obvious difficulty for this original model is the exact derivability, because it is impossible to derive the exact laws of T_P from T_F. To solve this problem, Kenneth F. Schaffner makes a revision that is called Generalized Nagel-Schaffner (GNS) model of reduction. The Schaffner proposal can be briefly formulated as follows: "T_F reduces T_P iff there is a corrected version T_P* of T_P such that, (a) T_P* is derivable from T_F given that the terms of T_P* are associated via bridge laws with terms of T_F, and (b) the relation between T_P* and T_P is one of, at least, strong analogy (sometimes also 'approximate equality', 'close agreement', or 'good approximation')" (Dizadji-Bahmani, F., Frigg, R., and Hartmann S., 2011, p. 398). The abbreviation DFH in this dissertation stands for Dizadji-Bahmani, F., Frigg, R., and Hartmann S. The

derivation of T_P* can be done in two steps, namely: (a) deriving a special version of T_F, called T_F*, by introducing auxiliary assumptions, and (b) replacing the relevant terms by their 'correspondents' using bridge laws to produce T_P* (DFH, 2011, p. 398). In this revised version bridge laws are crucial to this picture of reduction because "reduction is the deductive subsumption of a corrected version of T_P under T_F, where the deduction involves first deriving a restricted version, T_F*, of the reducing theory by introducing boundary conditions and auxiliary assumptions and then using bridge laws to obtain T_P* from T_F*". (DFH, S. 2011, p. 399) Also, this model can be represented as follows:

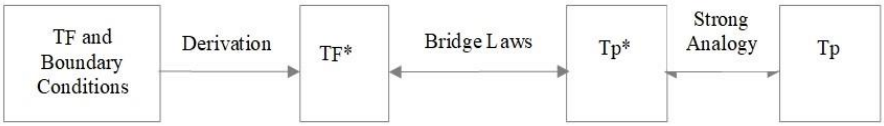


Figure 8.3. The generalized Nagel-Schaffner model of reduction (Source: DFH, S. 2011, p. 399)

DFH proposes several other improvements in their paper *Who's afraid of Nagelian Reduction?* as follows: (a) The first revision is about the status of bridge laws. According to DFH, p. 404, bridge laws cannot just be a mere convention but must be factual claims. There are two different kinds of bridge laws: The first kind is bridge laws that associate basic entities of T_P and T_F with each other. This kind is called entity association laws. The second kind is called property association laws. "They assert that the T_P-properties of a system stand in a relevant relation to the T_F-properties of that system, and the magnitudes of these properties stand in a relevant functional relationship" (DFH, p. 404). (b) The second revision is about the meaning or interest of reduction in the discussion about multiple realizations, which is a hot topic in the interdisciplinary context. Multiple realizations, it is said, undercut the explanatory power of reductions. Reductions are desirable for

two reasons: consistency and confirmation. Establishing a reductive relation between T_F and T_P ensures the consistency and co-tenability of both accounts. If we have two theories whose target domains are identical (or have significant overlap), we would expect evidence confirming one theory would also confirm the other theory. It can only happen if the two theories are connected, for example, in the reduction of thermodynamics to statistical mechanics. (c) The third revision is about two additional requirements for auxiliary assumptions: The first is the condition of non-redundancy, i.e., T_F must be used in the deduction of T_P*; that is, T_P* must not follow from the auxiliary assumptions alone. The second is the condition of immanence. The auxiliary assumptions must belong to the paradigm of T_F; i.e., auxiliary assumptions cannot be foreign to TF's conceptual apparatus. (d) The fourth revision is an additional condition for T_P* beside the five conditions formulated in Schaffner, 1967, p. 144: DFH, 2011 requires that T_P* must share with T_P all essential terms.

On the other hand, Van Riel proposed another revision of the current interpretation of Nagel's model of reduction, the official Nagelian model. According to Van Riel, the official Nagelian model has a tendency[1] to have less to do with observations. This tendency leads to some formal worries: "if reduction is a derivation plus (sometimes) bridge laws, then any theory would reduce to itself …; moreover, any theory would reduce to any inconsistent theory, and contrary to what one might expect, reduction is not an asymmetric relation" (van Riel, 2011, p. 354). In his reinterpretation, van Riel emphasizes ontological aspects of reduction – not only epistemological aspects. There are five points regarded as the main features of the "real

1 I employ here the word "tendency," because I do not think that all current thinkers despise the ontological or observational aspects of Nagel's model of reduction. For example, that DFH do not despise the observational aspects of Nagelian reduction. In the 2011 paper, their concern is to answer epistemologically the objections against especially Nagel's model of reduction and generally the idea of reduction, and they also attempt to correlate the phenomena explained by T_F and T_P to constructed T_F* and T_P* via Bayes' theorem.

Nagel model" in his reinterpretation, namely: "(1) Reduction is a relation holding among a great variety of scientific representational devices, among which theories play an important epistemological role. (2) Interesting reductions are explanations that consist in deductions that are carried out with the help of bridge laws, and they have to obey (some of) the relevant non-formal criteria (unification, appropriateness of reducing theory and bridge laws, and, if possible, correction should be involved in reduction). (3) Bridge laws are to be regarded as stating ontological links (identities or relations among extensions) in a posteriori. (4) The reduction is not direct (in the sense that it is not a case of theory explanation) – it goes together with explanations of the phenomena of the reduced theory by the reducing theory. (5) The Nagel's model is not a [mere] epistemological model of reduction" (van Riel, 2011, p. 371-2).

This GNS model of intertheoretical reduction with these improvements is very different from the structuralist model used here. However, we can still discuss how both models explain the phenomena of intertheoretical reduction in their own ways. Here I will discuss three points regarding this topic: The first point is the main differences between the structuralist theory of science and the GNS model. The second point is the difference in building the model of intertheoretical reduction. This point is related to the main GNS model of intertheoretical reduction. The third point is about the difference between the structuralist models and the GNS model related to the claim that the model must capture how the intertheoretical reduction should be related to observations. This last point is the main topic of all revisions of the GNS model above.

8.2.1.1. The Main Differences Between the GNS Model and the Structuralist Model

The main difference lies in the structuralist requirement that the scientific theories in question must be modeled in set theory to model their

inner logical structure. To build the model of a theory, we form the class of potential models $\mathbf{M_p}$ of the theories on the **T**-theoretical level and the class of partial potential model ($\mathbf{M_{pp}}$) on the **T**-non-theoretical level. An intertheoretical connection as a bridge between theories is a relationship between terms in $\mathbf{M_p}$ of the connected theories. This approach requires more complicated modeling than the GNS model but, at the same time, provides a more detailed and precise analysis. The differentiation between the **T**-theoretical level and the **T**-non-theoretical level in the structuralist model points out the form of the empirical claims (and intended applications) of the connected theories. It enables us to characterize the scope of the intertheoretical connections precisely, which represents the empirical objects or the concepts from other theory-elements. By characterizing the actual models of related theories, we can also characterize what kind of intertheoretical connections are there and how they connect the law or law-like statements of both theories. The second main difference is that, according to the structuralists, the intertheoretical reduction is more epistemological than ontological, although it should have some empirical basis – characterized by the partial potential models. It is related more to the structure of theories, rather than to reality itself. Finally, the third main difference is that, in the structuralist theory of science, the intertheoretical reduction is just one among several other intertheoretical connections (links), and the structuralists have already characterized some of them formally.

8.2.1.2. There Is No Generalized Structuralist Model of the Intertheoretical Reduction as Such

Whereas the main differences between the structuralist model and the GNS model can be found just by theoretically comparing both approaches, this dissertation exposes further differences by implementing the structuralist theory of science to model some intertheoretical

connections in some real areas of research in cognitive science. By modeling some of them, this dissertation uncovers a further difference between the structuralist model of intertheoretical reduction and the GNS model; namely, there is no such generalized model of intertheoretical reduction for the structuralists. Structuralists have characterized a formal definition of how a scientific theory can be reduced to another theory, but the structuralists have no intention of formalizing a general pattern of reduction for scientific practices. It is so because as we have seen in the previous chapters, the patterns of intertheoretical reduction can differ depending on the purposes and levels of explanation scientists have set for their research. In Chapters 5–7, three patterns of an intertheoretical reduction have been laid out – there could exist more. Chapter 5 lays out a model of intertheoretical reduction without modifications. A model of intertheoretical reduction with some modifications has been discussed in Chapter 6. Chapter 7 presents a model of reduction with a combination of several theories and modifications of previous theories.

A Reduction Model Without Modifications. The reduction model without modifications in Chapter 5 is exemplified by the intertheoretical reduction between the theory of forced compliance dissonance (DissF) as the reduced theory and the computational neuroscientific theory (CNT) as the reducing theory placed in the context of van Veen et al. 's research. In this case, CNT can immediately reduce DissF based on the assumption of the connection itself that on the brain's level the reducing dissonance between cognitions is in the *dorsal Anterior Cingulate Cortex (dACC)* represented by the communication network of neurons, which involve the neurons' activation and the connection's weight. We do not need to modify either CNT or DissF to any other theories before we build an intertheoretical reduction bridge/link between the concepts of both theories. The connected concepts of both theories as being modeled in the theory-elements **T**(DissF) and **T**(CNT) can be seen in Table 8.1.

Determining links	Set in **T**(DissF)	Set in **T**(CNT)	Note
λ_1	Time	T	This determining link connects the set of Time from the theory-element **T**(DissF) and the set of time (T) from the theory-element **T**(CNT). The connection is „is equal to" and therefore is bijective. This connection connects the points of time, when the dissonance or consonance happen, and the time, when the neurons in certain parts of the brain, such as the dorsal Anterior Cingulate Cortex (dACC), do the computational process that represents either dissonance or consonance.
λ_2	Cognition	The relation between the set of Activation (AV) and set of connection weight (CW)	This determining link connects the set *Cognition* in the theory element **T**(DissF) and the relation between the set of activation of neurons and the set of connection weight between neurons in **T**(CNT) according to the law-statement: For all $n_1,n_2 \in N$: $AV(n_1,t) = CW(n_2,n_1)$. $O(n_2)$ $I(n_1)$ $AV(n_1, t-1)$ (DIII-17(2)).
λ_3	Forcecom	The relation between Activation (AV) and connection weight (CW)	Since *Forcecom* is a subset of Cognition and this determining link connects the set *Forcecom* to the relation between the set of activation of neurons and the set of connection weights between neurons in **T**(CNT) according to the same law-statement, then λ_3 is a subset of λ_2.
λ_4	Disscog	The relation between the set of connection weights and itself	Since adjustment of cognition needs communication between neurons, the dissonance will raise intensive communication between neurons in certain parts of the brain, here is the dorsal Anterior Cingulate Cortex (dACC). The fourth determining link connects the *Discog* from **T**(DissF) to the relation between active connection weights through a certain period, when the dissonance happens. If the dissonance happens, the sum of the difference of active connections weight through time will be big. This phenomenon is confirmed by van Veen et al. 's research.
λ_5	Conscog	The relation between the set of connection weight and itself	The fifth determining link connects the *Conscog* from **T**(DissF) to the relation between active connection weights through a certain period, when the consonance happens. If the consonance happens then – dissonance does not happen, and the sum of the difference of active connections weight through time will be small.

Table 8.1. The Determining Links that Connect Concepts of **T**(DissF) and **T**(CNT)

In this case, the intertheoretical reduction brings us relatively little explanatory power about the cognitive (dissonance) processes and their relationship with the neural network in the human brain based on such common assumptions. The reduction delivers nothing more than a confirmation that cognitive dissonance processes are connected to neurons' activities in the *dorsal Anterior Cingulate Cortex (dACC)* of the brain; that is, there is high activity of neurons in the dissonance case and low activity of neurons in the non-dissonance cases. However, we cannot know, which neurons from the set of neurons of *dACC* can be connected to which cognition or cognition processes that have been observed. From the empirical result of the observation done by van Veen et al. by fMRI we can just confirm that such a dissonance reduction process happens and the intertheoretical reduction works.

A Reduction Model with Modifications. The second case and the third case are cases of reduction of the psychological theory to the artificial neural network in some simulations. By such simulations, cognitive scientists hope to understand how the neural network in our brains can perform a (computational) process such that the phenomena of cognition emerge. For that purpose, they use artificial neural networks, which are inspired by how neurons work. In this second and third cases, some modifications are needed such that the theories can be connected reasonably.

In this second case of reduction, Shultz and Lepper use the Hopfield network to simulate dissonance reduction. This simulation shows that the values of specific cognitions are changing in the process of dissonance reduction. To build this simulation, Shultz and Lepper implement the Hopfield Network with some modifications. The modifications include not only several additional conditions or assumptions but also adding some constants. They can be seen in Table 8.2:

No.	**T**(HN)	Modifications	**T**(HN for Consonant)
M_1	N	The first modification is creating two new subsets of *N*, namely N^+ and N^-	$N^+ \subseteq N$ $N^- \subseteq N$
M_2	C	The second modification is creating two new subsets of C, namely *Pairs* and *Conpairs*. *Pairs* is the relation between sets N^+ and N^-. *Conpairs* is a relation between *Pairs*.	Pairs $\subseteq N^+ \times N^-$, Pairs is a bijective relation. Pairs $\subset$ C Conpairs $\subseteq$ Pairs' Pairs C = Pairs $\cup$ Conpairs
M_3	ext_n	Removed	
M_4	out_n	Removed	
M_5		Adding a learning parameter called "$resist_i$."	$resist_i$
M_6	net_n	Modifying the law-statement.	$net_i = resist_i \Sigma_j w_{ij} a_j$
M_7		Adding a new set of points of time (*T*)	T
M_8		Adding a new set called *ceiling*, which is set to 1 for N^+ and is set to 0.5 for N^-	ceiling
M_9		Adding a new set called *floor*, which is set to 0 for all *N*.	floor
M_{10}	θ, act_n	Modification of the updating rule of activation of the neurons. Here the neurons are assumed just excitatory or inhibitory. Both concepts are reduced to A, where $a \in A$.	1) $a_i(t+1) = a_i(t) + net_i[ceiling - a_i(t)]$, when $net_i \geq 0$, 2) $a_i(t+1) = a_i(t) + net_i[a_i(t) - floor]$, when $net_i < 0$,
M_{11}	state	The term *state* is replaced by the term consonance	consonance
M_{12}	E	The term *E* is replaced by the term dissonance	dissonance
M_{13}		Addition of the *cap* parameter	cap
M_{14}		Addition of the *rand%* parameter	rand%

Table 8.2. Modifications of the Hopfield Network for the Consonance Model

With these modifications to the Hopfield network's theory-element, we defined a new theory-element of the Hopfield Network that is built especially for the Consonant model of simulation (**T**(HN for Consonant)). Therefore, the intertheoretical reduction is actually between **T**(DissB) as the reduced theory and **T**(HN for consonant) as the reducing theory. The determining links for this intertheoretical reduction can be shown in Table 8.3:

Determining links	**T**(DissB)	**T**(HN for Consonant)	Explanation
λ_1	Cognitions	Pairs	λ_1 connects the set *Cognition* to the set *Pairs* and λ_1 is bijective.
λ_2	Disscog	Conpairs	λ_2 connects the set *Disscog* to the set *Conspairs* and λ_2 is bijective.
λ_3	Conscog	Conpairs	λ_3 connects the set *Conscog* to the set *Conspairs* and λ_3 is bijective.
λ_4	Pairdiss	W	λ_4 connects the set *pairdiss* to the set of connection weight *W*, and λ_4 is bijective.
λ_5	Paircons	W	λ_5 connects the set *paircons* to the set of connection weight *W*, and λ_5 is bijective.
λ_6	Pairimp	imp	λ_6 connects the set *Pairimp* to the set *imp* and λ_6 is bijective.
λ_7	Diss	Dissonance	λ_7 connects the set *diss* to the set *Dissonance* and λ_7 is bijective.

Table 8.3. The Determining Links that Connect Concepts of **T**(DissB) and **T**(HN for Consonance)

A Reduction Model with a Combination of Several Theories and Modifications. In the third case, the connectionist model uses the two-layers feed-forward neural network with the Rosenblatt perceptron and the delta rule as a learning rule for this network. In this case, we see that for specific intertheoretical reduction the successfully reducing theory is not one single original theory-element, but a combination (and even with modification) of several theory-elements connected by determining links. The structuralist metatheory requires us to model this combination to analyze the intertheoretical reduction accurately. The combination of these theories can be modeled by implementing the V-pattern and the strategy above. The combination of these theory-elements is done by the determining links, which help us to build the unifying model as presented in Table 8.4:

No.	**T**(RP)	**T**(2L-FFNN)	**T**(DL)	Explanations of Determining links	**T**(RP+2L-FFN+DL)
1	N	N’		The first link determines that *N* be identical with *N′* because both are the set of all neurons in the networks. It can be simplified by eliminating the redundant set *N′* in the unifying model.	N
2	N/N_0	N_{out}	N*	The second link determines that all *N* besides N_0 be identical with N_{out} and *N**. Therefore, for the reason of simplification, we include only the N_{out} in the unifying model.	N_{out}
3	N_0	N_{in}		The third determining link equates N_0 and N_{in} because both sets are the sets of input-neurons. The unifying model includes only N_{in}.	N_{in}
4	C	C’	C*	The fourth determining link connects all sets of connections in the three theory-elements because they refer to the same object in the application. Therefore, we can use just *C*.	C
5	B		B*	The fifth determining link identifies the set *B* from **T**(RP) as identical with *B** from **T**(DR). Hence, the unifying model needs just *B*.	B
6	W	W’		The sixth determining links equate *W* and *W′*. Thus, we need just *W* in the new unifying model.	W
7	W, W_0		W*	The seventh determining link defines that the connection weight **W*** of **T**(DR) is a unification of both *W* and W_0 of **T**(RP). Hence, we still need W_0.	W_0
8	Inp	ext_n	Inp*	The eighth determining link equates *Inp, ext_n, Inp** as the network-inputs for the unifying model. Therefore, we need only one of them.	Inp
9	Outp	out_n	OUTn	The ninth determining link equates *Outp, outn,* and *OUTn* because all of them is the set of the actual output of the network. We use only Out_n	out_n
10	fnet	net_n		The tenth determining link connects net_n to *fnet* as its result.	fnet, net_n
11	fact	act_n		The eleventh determining link connects act_n to *fact* as its result.	fact, act_n
12	fout	out_n		The twelfth determining link connects out_n to *fout* as its result.	fout

Table 8.4. The Determining Links that Connect Concepts of **T**(RP), **T**(2L-FFNN), and **T**(DL)

We can use this unifying model to simulate several phenomena of dissonance reduction according to the connectionist model. However, van Overwalle and Jordens used it in their research for a specific case, namely, forced compliance dissonance among children when they are exposed to a toy and punishment. To build this simulation, the unifying theory-element is adjusted as follows: (1) The unifying model consists of 4 neurons in two layers. (2) The connectionist model implements the linear activations for the output perceptron(s) in the neural network. (3) The activation of an input-neuron will be set 1 or 0 for a toy, representing "present" or "not present." Moreover, for the threat, an input neuron will be set 0, +0.5, or +1 to represent "not present," "mild threat," or "severe threat." (4) The expected activation of output-neurons will be set of +1 when the outcome is present (e.g., play or happy), zero when absent (e.g., not play, or moderate affect), and -1 when the opposite outcome is present (e.g., unhappy). (5) Also, the learning rate for this simulation is set between 0 and 1.

The connectionist simulation requires not only the combination and modification of the several theory-elements in the artificial neural networks but also some modifications of the theory-element of forced compliance dissonance **T**(DissF) itself. The modifications are listed in Table 8.5:

No.	**T**(DissF)	Modifications	**T**(DissF for Connectionist)
M_1	Cognition	We build a subset of *Cognition* that is called *thought*.	thought $\subseteq$ Cognition
M_2	Cognition	We build another subset of *Cognition* that is called *behavior*.	behavior $\subseteq$ Cognition
M_3	Cognition	We build another subset of *Cognition* that is called *emotion*.	emotion $\subseteq$ Cognition
M_4		We build a new set called *attitude*, which expresses the relation between *thought* and *behavior* or between *thought* and *emotion*. This new set also should be the superset of *Disscog* and *Conscog*	attitude $\subseteq$ thought 'behavior $\vee$ thought 'emotion. Disscog $\subset$ attitude Conscog $\subset$ attitude Disscog $\cup$ Conscog $\subseteq$ attitude
M_5		We add a new magnitude of the intensity of attitude (*attint*), which expresses a function that maps every element of *attitude* to a rational number to express the strength of the attitude.	attint: attitude $\rightarrow$ IR_0
M_6	Forcecom	We set the set *Forcecom* as a subset of *thought*	Forcecom $\subseteq$ thought
M_7		We add a new subset of *attitude*, called *subattitude*, which consists of the relation between *Forcecom* and *behavior* or between *Forcecom* and *emotion*.	subattitude $\subseteq$ Forcecom$\times$behavior $\vee$ Forcecom$\times$emotion subattitude $\subseteq$ attitude
M_8		We add a new subclass of the intensity of attitude, called *subattint*, which expresses a function that maps every element of *subattitude* to a rational number to express their strength.	subattint: subattitude $\rightarrow$ IR_0 subattint $\subseteq$ attint

Table 8.5. Modifications of the Theory Forced Compliance Dissonance for the Connectionist Model

These modifications in DissF are needed because the simulation is intended to reflect "a view of the mind as an adaptive learning mechanism, where cognitive dissonance is seen as a relatively rational process in which people seek causal answers for why they think, feel or behave inconsistently" (van Overwalle and Jordens, 2002, p. 205). To make the simulation works several concepts in both modified theories have to be connected by the following determining links in Table 8.6:

Determining links	**T**(DissF for Connectionist)	**T**(RP+2L-FFN+DL for Connectionist)	Explanation
λ_1	Cognition	N	λ_1 is a determining link that connects the set *Cognition* to the set of Neurons (N) and is a bijective relation.
λ_2	Thought	N_{in}	λ_2 is a determining link that connects the set *thought* to the set of Input Neurons (N_{in}) and is a bijective relation.
λ_3	Forcecom	N_{in}	λ_3 is a determining link that connects the set *Forcecom* to the set of Input Neurons (N_{in}) and is a bijective relation.
λ_4	Behavior	N_{out}	λ_4 is a determining link that connects the set *behavior* to the set of Output Neurons (N_{out}) and is a bijective relation.
λ_5	Emotion	N_{out}	λ_5 is a determining link that connects the set *emotion* to the set of Output Neurons (N_{out}) and is a bijective relation.
λ_6	Attitude	C	λ_6 is a determining link that connects the set *attitude* to the set of connection (C) and is a bijective relation.
λ_7	Subattitude	C	λ_7 is a determining link that connects the set of *subattitude* to the set of connection (C) and is a bijective relation.
λ_8	Attint	W	λ_8 is a determining link that connects the set of *attint* to the set of connection weight (W) and is a bijective relation.
λ_9	Subattint	W	λ_9 is a determining link that connects the set of *subattint* to the set of connection weight (W) and is a bijective relation.
λ_{10}	Imp	Inp	λ_{10} is a determining link that connects the set of important of cognition (*imp*) to the set of input (*Inp*) and is a bijective relation.
λ_{11}	Reward	Inp	λ_{11} is a determining link that connects the set *reward* to the set of input (*Inp*) and is a bijective relation.
λ_{12}	Behavior	Out	λ_{12} is a determining link that connects the set *behavior* to the set of desired output (*Out*) and is a bijective relation.
λ_{13}	Emotion	Out	λ_{13} is a determining link that connects the set *emotion* to the set of desired output (*Out*) and is a bijective relation.
λ_{14}	Diss	Error	λ_{14} is a determining link that connects the set of dissonance (*diss*) to the set *Error* and is a bijective relation.

Table 8.6. The Determining Links that Connect Concepts of **T**(DissF for Connectionist) and **T**(RP+2L-FFN+DL for Connectionist)

A Reduction Model of Two Theories where the Newcomer Becomes a Generalization the Old One. The model of intertheoretical connection of the McCulloch-Pitts neuron and the Rosenblatt perceptron is a model of such intertheoretical reduction. This model shows how two theories developed in two different disciplines can relate to each other because one of them generalize the other. In this case, the Rosenblatt perceptron, which came later and was developed in artificial intelligence, is a further development of the McCulloch-Pitts neuron, which came first and was developed in neuroscience. Synchronically, the intertheoretical connection of both theories is an intertheoretical reduction where the Rosenblatt perceptron reduces the McCulloch-Pitts neuron. The determining links which play a decisive role in the intertheoretical reduction, or respectively the intertheoretical specialization are listed in Table 8.7:

Determining links	Set in **T**(RP)	Set in **T**(MCP-N)	Note
λ_5	$B \times W_0$	θ	a determining link that connects the relation between the bias of the neurons (*B*) and its connection weight (W_0) in the potential model of the Rosenblatt perceptron (x) to the set of a threshold of each neuron (θ) in the potential model of the McCulloch-Pitts Neuron (x'). λ_5 is surjective.
λ_6	Inp	Inp'	a determining link that connects the set of inputs (*Inp*) in the potential model of the Rosenblatt perceptron (x) and the set of inputs (*Inp'*) in the potential model of the McCulloch-Pitts Neuron (x'). λ_6 is surjective because Inp $\supset$ Inp'.
λ_7	Outp	Outp'	a determining link that connects the set of outputs (*Outp*) in the potential model of the Rosenblatt perceptron (x) and the set of outputs (*Outp'*) in the potential model of the McCulloch-Pitts Neuron (x'). λ_7 is surjective because Outp $\supset$ Outp'.
λ_9	fact	fact'	a determining link that connects the activation function of neurons (*fact*) in the potential model of the Rosenblatt perceptron (x) and the activation function of neurons (*fact'*) in the potential model of the McCulloch-Pitts Neuron (x'). λ_9 is surjective because fact $\supset$ fact'.
λ_{10}	fout	fout'	a determining link that connects the output function of neurons (*fout*) in the potential model of the Rosenblatt perceptron (x) and the output function of neurons (*fout'*) in the potential model of the McCulloch-Pitts Neuron (x'). λ_{10} is surjective because fout $\supset$ fout'.

Table 8.7. The Determining Links that Connect Concepts of **T**(RP) and **T**(MCP-N)

8.2.1.3. The Empirical Status of the Intertheoretical Reduction

The third point to mention concerns the epistemological and ontological status of the intertheoretical connections, especially intertheoretical reduction. These issues are connected to some revisions made by DFH and by van Riel for the GNS model above. In the structuralist model, we can explain the epistemological status of an intertheoretical connection in general by the links connecting the potential models of both

connected theories. The structuralist model uses the interpreting links to show that the intertheoretical reduction has something to do with the explained phenomena. The interpreting links connect the potential models of one theory-element to the partial potential model of the other theory-element.

To understand how this works, we should remember what kinds of elements of the potential models and the partial potential models of a theory-element. A potential model consists of the sets that represent all basic concepts of the theory and all the basic relations among those concepts. Most of these concepts and relations are **T**-non-theoretical, but some of them are **T**-theoretical. By defining the partial potential models $\mathbf{M_{pp}(T)}$, we omit the **T**-theoretical elements from the potential models $\mathbf{M_p(T)}$ of a theory-element **T** by using a function r: $\mathbf{M_p(T)} \rightarrow \mathbf{M_{pp}(T)}$. Now we have the partial potential models of **T** that only consist of **T**-non-theoretical elements. Suppose we have a reducing theory **T*** that is connected partially through the intertheoretical reduction to **T** as the reduced theory. The intertheoretical reduction is a set that consists of several determining links, which connect several concepts or relations from the $\mathbf{M_p}$ of both theories. We can define the echelon subsets e of the $\mathbf{M_p}$ of both theories, consisting of only and all the connected concepts or relations related by the intertheoretical reduction. The intertheoretical reduction is now in the form of entailment links that connect both echelon subsets. To determine the empirical claim of the intertheoretical reduction, we can project $e_1(\mathbf{T})$ to the field of partial potential models $\mathbf{M_{pp}(T)}$ by using a function r*: $e_1(\mathbf{T}) \rightarrow f_1(\mathbf{T})$ and get $f_1(\mathbf{T})$ as an echelon subset of $\mathbf{M_{pp}(T)}$. $f_1(\mathbf{T})$ is the set of local empirical claims of the intertheoretical reduction on the side of **T**. The interpreting links connecting the $\mathbf{M_{pp}(T)}$ and the $\mathbf{M_p(T^*)}$ show us, which non-theoretical concepts of **T**, which come from other theories, are related by this intertheoretical reduction to **T***; they are defined by $f_1(\mathbf{T})$.

The **T**-non-theoretical elements that can be defined through defining f_1(**T**) – where **T** consists of the reduced theory-element – in the models defined in Chapters 5–7 can be shown as follows: In the first case of Chapter 5, the forced compliance dissonance is reduced by the computational neuroscientific theory. The structuralist model of intertheoretical reduction leads to characterize the **T**-non-theoretical terms of the forced compliance dissonance in Table 8.8:

no.	The **T**-non-theoretical Terms (Concepts)	Category: Observational term or concept of other theories	Explanation
1	Time	an observational term	a set of points of time.
2	Cognition	an observational term	a set of cognitions.
3	Disscog	an observational term	a set of pairs of cognitions, which are dissonant each other.
4	Conscog	an observational term	a set of pairs of cognitions, which are consonant with each other.
5	Forcecom	an observational term	a set of forced compliance, which is a subset of Cognition.

Table 8.8. The Local Empirical Claims of the Intertheoretical Reduction between **T**(DissF) and **T**(CNT)

In the second case in Chapter 6, namely consonance simulation, the theory of cognitive dissonance is reduced by the Hopfield network. The structuralist model of intertheoretical reduction characterizes the **T**-non-theoretical terms of the theory of cognitive dissonance in Table 8.9:

no.	The **T**-non-theoretical Terms (Concepts)	Category: Observational term or concept of other theories	Explanation
1	Cognition	An observational term	The set of cognitions
2	Disscog	An observational term	The set of pairs of cognitions, which are dissonant each other
3	Conscog	An observational term	The set of pairs of cognitions, which are consonant with each other

Table 8.9. The Local Empirical Claims of the Intertheoretical Reduction between **T**(DissB) and **T**(HN for Consonance)

In the third case in Chapter 7, the connectionist simulation, the modified theory of forced compliance dissonance is reduced to the two-layer neural network, which consists of three theories, i.e., the Rosenblatt perceptron, the two-layer architecture of the feed-forward neural network, and the delta rule as its learning rule. The **T**-non-theoretical terms of the modified forced compliance dissonance characterized by the modeling in Chapter 7 are in Table 8.10:

no.	The **T**-non-theoretical Terms (Concepts)	Observational term or concept of other theories	Explanation
1	Cognition	An observational term	The set of cognitions
2	thought	An observational term, but also a concept from another theory	The set *thought* is a subset of the set C*ognition* concerning the attributional reformulation advocated by Cooper and Fazio (1984)
3	behavior	An observational term, but also a concept from another theory	The set *behavior* is a subset of the set *Cognition* concerning the attributional reformulation advocated by Cooper and Fazio (1984)
4	emotion	An observational term, but also a concept from another theory	The set *emotion* is a subset of the set *Cognition* concerning the attributional reformulation advocated by Cooper and Fazio (1984)
5	attitude	An observational term, but also a concept from another theory	The set *attitude*, as a superset, unites both the set of cognitive dissonance (*Disscog*) and the set of cognitive consonance (*Conscog*), which are understood as attitude to objects
6	Forcecom	Observational term	The set of forced compliance, which is a subset of *Cognition*.
7	subattitude	An observational term, but also a concept from another theory	The subset *subattitude* is a subset of the set attitude concerning the set *Forcecom*
8	imp	Observational term	The set of importance of cognition
9	reward	Observational term	The set of magnitude of reward or punishment

Table 8.10. The Local Empirical Claims of the Intertheoretical Reduction between **T**(DissF for Connectionist) and **T**(RP+2L-FFN+DL for Connectionist)

In the third case, we can see the reasons for the distinction between **T**-theoretical and **T**-non-theoretical concepts, namely that the observational terms are characterized not merely by a pure observation alone, but also by involving other theories. It is not decided whether the terms, which we categorize as "observational terms," are purely observational terms or employ concepts from other theories. The limitations of this dissertation do not allow me to be more explicit about this issue.

The fourth case, namely the intertheoretical reduction of the Rosenblatt perceptron to the McCulloch-Pitts neuron, is not a case of the global intertheoretical relation but a case of the local intertheoretical connection since both theories are close together though from two different disciplines. The McCulloch-Pitts neuron becomes a special case of the Rosenblatt perceptron. Therefore, we do not need to discuss a local intended application. The Rosenblatt perceptron's empirical claims become smaller in its specialization, the McCulloch-Pitts neuron because of an additional set T and the restrictions applied to the sets θ, Inp, and Outp.

After discussing four cases above, some conclusions about the structuralist model's distinctive characteristics in comparison with the GNF model of intertheoretical reduction can be drawn: (1) The structuralist is required to model the theories and the intertheoretical reduction in set theory or model theory. This requirement allows for a higher degree of distinctness and sharpness of analysis. (2) The structuralist model does not build a general model of intertheoretical reduction like the GNS model. The structuralist equips us with definitions of theory-element and various intertheoretical connections and relations – and reduction is just one of them – to model and analyze those relationships case by case with great detail and accuracy. (3) By implementing the r* function, the structuralist can also specify how its model of intertheoretical reduction refers to the observational fields. It means that the intertheoretical reduction is not merely epistemological, and it does have a certain relation with the observed phenomena. In the structuralist theory of science, there are various types of intertheoretical connections that work together to produce a successful intertheoretical reduction. These primary intertheoretical connections are determining links and entailment links that operate at the **T**-theoretical level. Both determine whether the relation being constructed – not only reduction – is full or partial. Another type of connection is the interpreting links,

which operate at the **T**-non-theoretical level and define the reduction's empirical claims.

8.2.2. Unity of Science

In the discussion about the unity of science, there are two opposite positions. The first one is the position of philosophers who believe in the notion of the unity of science. They have a long history of how they try to find a basis for such a notion. If most of the history of philosophy – since antiquity until modern philosophy – the unity of science or knowledge had a speculative or metaphysical basis, the first time the idea of the unity of science got a more empirical basis, was when some philosophers in the Vienna Circle began their movement following Mach and pursuing the following goals: "1. To create new foundations for physics with strong consideration of the results of sense physiology; one could even speak of the attempt to give the concepts and principles of physics a psychophysiological basis. 2. Restore the unity of all empirical sciences. 3. To finally 'eradicate' the metaphysical speculations from the field of science." (Moulines, 2008, p. 26). This current is still being developed and refined until now. The other position is that of the Stanford philosophers, who do not believe in the idea of the unity of science. They are John Dupré, Ian Hacking, Peter Galison, Patrick Suppes, and Nancy Cartwright. They assume the notion of the disunity of science and plurality based not only on the methodological point of view toward science and scientific practice but also on a metaphysical point of view. (Cat, 2017) Dupré has characterized this position by three pluralistic theses as follows: "(1) against essentialism, there is always a plurality of classifications of reality into kinds; (2) against reductionism, there exists equal reality and causal efficacy of systems at different levels of description, that is, the micro level is not causally complete, leaving room for downward causation; and (3) against epistemological monism, there is no single methodology that supports a single criterion of scientificity, nor a

universal domain of its applicability, only a plurality of epistemic and non-epistemic virtues" (Cat, 2017).

Between both positions, this dissertation takes a unique position. On the one hand, it takes a view similar to the disunity of science. The structuralist metatheory of science starts with an assumption that scientific theories are already there, created through various assumptions, approaches, goals, etc., without supporting the agendas of essentialism, reductionism, and epistemological monism. In the modeling, in Chapters 3, 5–7, we can see that our models do not make any claims about those three notions, which are generally the basis of the unity of science. However, it does not mean that the structuralist theory of science abandons the notion of the unity of science. It also supports a certain notion of the unity of science, but not in the fashion above. This position has a great similarity with the integrative pluralism (Mitchell, 2003).[2]

The structuralist metatheory of science sees that being in connection with other scientific theories is one of the essential features of a scientific theory. In Chapter 2, many kinds of intertheoretical connections have been laid out to describe how scientific theories stay connected with each other in a synchronic relationship. In *Intertheoretical Relations and the Dynamic of Science* (2014), Moulines explicates several other intertheoretical connections in diachronic perspective. According to the structuralist metatheory of science all these intertheoretical connections build theory-nets and theory-holons as their results both in synchronic and diachronic perspectives. According to the structuralist theory of science, the unity of science, at least tendentially, is based on intertheoretical connections that connect the classes or concepts of the connected theories (for some practical reasons). Let us see how our investigation has provided some ground for ascertaining this tendency toward a unification of science, at least partially.

[2] A comparison and a relation between the structuralist theory of science and the integrative pluralism will not be discussed here because they are not the focus of this dissertation.

First, in Chapters 5–7, theories from the same or various disciplines are connected and build (fragments of) a theory-holon. These connections can be modeled and precisely characterized by implementing model theory. In this way, certain concepts between two or more theories can be connected precisely. This strategy works very well to identify the intertheoretical connections within a discipline and an interdisciplinary setting.

Second, in Chapters 5–7, the intertheoretical relations modeled are based on several real cases. The scientific theories are not automatically connected by themselves but are connected through the scientists' activities for specific practical purposes or goals, which have been predetermined before. Connecting two concepts from two theories, mainly from different disciplines, is not a simple task. There are some cases where the connections can be identified clearly, such as in the Festinger theory of cognitive dissonance and its specializations or the relation between the McCulloch-Pitts Neuron and the Rosenblatt perceptron. However, sometimes the connections are not as clear as in the case of the forced compliance dissonance and the computational neuroscientific theory. There are cases where we need to combine several theories in order to be clear enough about the existing connections, such as in the intertheoretical connections between the forced compliance dissonance and the unified theory coming from the Rosenblatt Perceptron, the two-layers feed-forward neural network, and the delta rule. Sometimes we also need to determine several constants and modify the theory so that the connections ‘work’ well, such as in the relation between the theory of cognitive dissonance and the Hopfield network. This dissertation demonstrates that the complexity of the problem of intertheoretical connections increases when we deal with interdisciplinary relations.

Third, in the simulation case, we see that a simulation must have a more specific goal than an explanation because it must cover additional aspects of the phenomena simulated. To model intertheoretical connections

between a theory of specific phenomena and the computational theories for simulation, we have several things to do, namely (1) modification of the simulated theory to make it suitable to the simulation's goal. the modification can be seen as a specialization of the original theory-element. (2) Building a simulation combination and modification of several simulating theories – normally computational theories – are needed by determining certain constants or constraints, such that the simulation becomes realistic enough. (3) A complete model of intertheoretical connections for a simulation can only be modeled after finishing both steps above.

In connection to the idea of the unity of science, this research shows that the structuralist theory of science is powerful enough to model various intertheoretical connections in real science, not only within one single discipline but also between several disciplines. By modeling and analyzing those intertheoretical connections, my dissertation shows that the unity of science without essentialism, reductionism, and epistemological monism is possible. The prospect of the unity of science responds to the possibility of different kinds of intertheoretical connections connecting various scientific theories from different fields, thereby forming (fragments of) theory-holon. It corresponds to the contention that being connected with other theories is one of the essential characteristics of real scientific theories.

8.3. Interdisciplinary Research and Cognitive Science

As discussed in Chapter 1 the new epoch of interdisciplinarity began with uneasiness about the loss of the unity of science. Since it is a recent trend in science, people still are worried about how to understand interdisciplinarity precisely. It is not an easy task, because there are many similar-sounding terms around. The differences between them were explained in Chapter 1. For understanding and practicing interdisciplinarity on the level of theoretical integration, it is not enough to characterize the

meaning of those terms precisely but also the connections between them that produce realistic and fruitful interdisciplinary researches. In Chapters 5–7, this dissertation discussed three models of intertheoretical relations of three distinctive interdisciplinary researches from the field of cognitive science. These cases show how the structuralist theory of science lets us map the intertheoretical connections between theories from various disciplines within real research programs.

The first is the intertheoretical reduction between Festinger's theory of cognitive dissonance from psychology and the Hawkins-Kandel Computational Neuroscientific Theory (CNT). This case is an example of the case of the mind-body problem: How our psychological phenomena are related to how the human brain works. In this case, the research examined a relation between the phenomena of cognitive dissonance and the activity of neurons in the brain. In this modeling, the theory of forced compliance dissonance, which underlies the van Veen et al. psychological experiment, is connected to the Hawkins-Kandel computational neuroscientific theory which can be applied to explain the activity of neurons of the *dorsal Anterior Cingulate Cortex (dACC)* during the moments of dissonance. The structuralist modeling of the intertheoretical relation serves as tools to specify the relations among terms from both theories according to this research.

The second case is a case of analogy between mind and computer, especially artificial intelligence. Shultz and Lepper built a simulation of the Festinger theory of cognitive dissonance by using the Hopfield network. This simulation is called the consonance model. This simulation aims to represent the mind as a mechanism that maintains some equilibrium and the dissonance reduction as a solving of the constraint satisfaction problem between someone's beliefs and behaviors. The structuralist modeling of the intertheoretical connections is applied to map not only the connection

among the terms of both theories but also the modifications of the Hopfield Network to meet the requirements and the goal of the simulations.

Finally, the third case, the case of connectionist simulation, is also a case of simulation of cognitive dissonance by using artificial intelligence. Nevertheless, the goal of the simulation and the type of artificial neural network are different. The purpose of the connectionist model is to simulate "a view of the mind as an adaptive learning mechanism, where cognitive dissonance is seen as a relatively rational process in which people seek causal answers for why they think, feel or behave inconsistently" (van Overwalle and Jordens, 2002, p. 205). The connectionist model implements a combination of the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule. This combination of the three and the forced compliance dissonance theory itself still need to be adjusted according the purpose of the simulation, before they can be combined. The detailed steps for building the structuralist model for this simulation are explained in Chapter 7.

Both cases in Chapters 6 and 7 show that to simulate certain phenomena, just applying some theories is not enough. A particular goal of simulation has to be set for determining the combination and adjustments to the theories. It leads to an interesting result, that about specific phenomena it is possible to build various simulations by setting different goal, emphasizing and focusing on different aspects, applying and combining different theories. Various simulations made our understanding of the phenomena more diverse in perspective. However, it brings us a new challenge to integrating them, such that diverse knowledge of certain phenomena does bring to comprehensiveness and not to contradiction. For answering this challenge, the author believes that the structuralist theory of science can be applied as follows: (1) The first step is characterizing the unifying theory-element for each simulation by implementing V-pattern and strategy. (2) The second step is adjusting and combining the unifying

theory-elements to build a unifying theory-element for unifying simulation. Of course, we still need to make further research to test how it works.

As we have discussed in Chapter 1, Ezrquerro and Manrique categorized the intertheoretical relations in cognitive science in the following: (1) classical view, (2) connectionist revision, (3) pragmatist approach, and (4) the reductionist approach. Different from the structuralist model of intertheoretical relation, this categorization is based on different positions regarding the notion of a privileged level. In connection with these kinds of intertheoretical relation in cognitive science, the structuralist model does not belong to one of them and does not serve their agendas. However, the structuralist model of intertheoretical connection can be implemented by them to build their model formally. The structuralist modeling offers detailed modeling and precise analysis in return. This offer can also be seen in the work of John Bickle, who applied the structuralist model for the reductionist approach, and in this dissertation, which applies to connectionist and other reductionist approaches.

Chapter 9
Some Concluding Remarks and Prospects for Future Research

The structuralist theory of science is a fruitful theory to model single scientific theories so that we can understand their inner structure and model the intertheoretical connections between some theories within one and the same discipline and between theories from various disciplines.

For modeling intertheoretical connections, the structuralist theory of science uses two most basic intertheoretical connections, namely determining links and entailment links. From these basic intertheoretical connections, we can model the various kinds of intertheoretical relations, either diachronically or synchronically. As for the results of intertheoretical connections, the structuralist theory of science differentiates them into two types, namely theory-net and theory-holon. Theory-nets are results of intertheoretical connections between theories in a close relationship, whereas theory-holons are results of intertheoretical connections between theories in global science. They can be either from one discipline or various disciplines. However, the paradoxical result from the analysis of empirical claims of theory-holons is that their empirical claims are not global, but local.

In this dissertation, the structuralist theory of science is applied to model and analyze intertheoretical connections between the theories from a single discipline and from various disciplines – interdisciplinary cases – in real scientific practice. There are four models of intertheoretical relations between theories from a single discipline presented here. They are (1) the model of intertheoretical specialization between the forced compliance dissonance and the general theory of cognitive dissonance made by

Westermann, (2) the model of intertheoretical specialization of architectures of neural networks, and (3) the unifying relation between the Rosenblatt perceptron, the two-layers feed-forward neural network, and the delta rule. The model of intertheoretical relations for interdisciplinary cases presented here are (1) The Festinger theory of cognitive dissonance as related to the Hawkins-Kandel computational neuroscientific theory (CNT), (2) the model of intertheoretical reduction and intertheoretical specialization between the McCulloch-Pitts neuron model and the Rosenblatt perceptron – It is a unique relationship because both theories are from two different fields, but they are very closely related, (3) the Festinger theory of cognitive dissonance as related to the Hopfield network for the consonance model of simulation, and (4) the Festinger theory of cognitive dissonance as associated with the unified theory between the Rosenblatt perceptron, the two-layers feed-forward neural network and the delta rule for the connectionist model of simulation.

Based on all models that have been built in this dissertation, some conclusions can be drawn. First, the intertheoretical relations in interdisciplinary fields can be modeled formally like other intertheoretical relations by connecting the theories' potential models. The main difference lies in the basic sets of the respective potential models, whose elements depend on the discipline, to which the theories belong. Second, the most (if not all) models of intertheoretical connections in interdisciplinary research will not fulfill the definition of entailment links concerning the respective potential model. The determining links play crucial roles in building their models. In most (if not all) cases, entailment links connect only the echelon partial subset of the respective partial models. Third, the case of synchronic intertheoretical relation between the McCulloch-Pitts neuron and the Rosenblatt perceptron is not an interdisciplinary relation, but cross-disciplinary relation, where a theory of neuron in neuroscience is taken over to formulate a theory of perceptron in computer science (see Chapter 1,

footnote 1, p. 8). For cross-disciplinary relations, it is safe for now to consider the first point is correct.

This research delivers some contributions to the philosophy of science as well as to interdisciplinary studies, especially in cognitive science, from where the examples come. For the philosophy of science, this research shows that there are many kinds of intertheoretical connections – not only intertheoretical reduction – that deserve more attention from the scientists and the philosopher of science. Secondly, to model a successful intertheoretical reduction, we need other kinds of intertheoretical connections that map the links between the terms or concepts of the connected theories. Thus, the structuralist theory of science can deliver a more detailed model than the generalized Nagel-Schaffner model. Thirdly, this research shows us a unique kind of the unity of science. The idea of the unity of science envisaged in the present work is not based on essentialism, reductionism, and methodological monism. However, it is based on practical reasons for and the goals of the researches. The unity of science promoted here is the result of connecting scientific theories of single or various discipline(s) through intertheoretical connections. However, the unity of science envisaged here does not only contain the epistemologically correct links without observational or empirical truth.

Moreover, for interdisciplinary research, primarily cognitive science, this research is the first attempt to deliver mathematical models of intertheoretical connections between theories. We still have to build more models or maps of intertheoretical links for more interdisciplinary research to obtain a more comprehensive explanation of how scientific theories can be combined in interdisciplinary research to achieve its goal effectively. From this attempt to build several models of intertheoretical connections, this dissertation characterizes the V-pattern and strategy, that can serve as a procedure for combining several theories given as theory-elements and for building a new unifying theory-element. Finally, to explain some

phenomena – of which we cannot have a direct observation – some scientists attempt to develop a simulation to deeper comprehend the phenomena in question. It is the case for both simulations that we model in Chapters 6 and 7. To create a simulation, we need a more complex combination of theories rather than to deliver an explanation because we attempt to mimic the phenomenon itself. The structuralist modeling can model not only intertheoretical connections for creating a simulation but also specific adjustments needed.

Indeed, this research is limited to modeling and mapping the intertheoretical relations both within a discipline and in some interdisciplinary contexts. Still, this research can be extended through several possible types of research. For example, (1) Inspired by the application of Bayesian networks to the general Nagel-Schaffner account of reduction, we can apply Bayesian networks to the structuralist account to measure the degree of confirmation of the intertheoretical connections. (2) With the development of machine learning and artificial intelligence, I see a possibility to combine this research with some approaches in these fields such as the artificial neural network and reinforcement learning to write a computer program that might help us in combining theories or other intelligent models, tracking, and documenting the relations between them.

Bibliography

Anderson, James A (2006): *An Introduction to Neural Networks 2nd Ed*. Cambridge: MIT Press.

Baker, Lynne R. (1999): Folk Psychology in Wilson, Robert A., and Keil, Frank C. (1999): *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, Massachusetts: MIT Press, pp. 319-320.

Bartels, Andreas, Stöckler, Manfred (Eds.) (2007): *Wissenschaftstheorie. Ein Studienbuch*. Paderborn: Mentis.

Bartelborth, Thomas (1996): *Begründungsstrategien: Ein Weg durch die analytische Erkenntnistheorie*. Berlin.

Bartelborth, Thomas (2002): Explanatory Unification. In *Synthese* 130, 2002, pp. 91-107.

Balzer, Wolfgang (1982): *Empirische Theorien: Modelle Strukturen Beispiele*. Braunschweig: Springer.

Balzer, Wolfgang (2009): *Die Wissenschaft und ihre Methoden: Gründsätze der Wissenschaftstheorie*. München: Verlag Karl Alber.

Balzer, Wolfgang, and Moulines, C. Ulises. (Hg.) (1996): *Structuralist Theory of Science: Focal Issues, New Results*, in Perspectives in Analytical Philosophy, Bd. 6. Berlin: de Gruyter.

Balzer, Wolfgang, Moulines, C Ulises, and Sneed, Joseph D., (1986): *The Structure of Empirical Science: Local and Global*, in Barcan, Marcus et al., (Eds.) (1986): *Proceedings of the 7th International Congress of Logic, Methodology, and Philosophy of Science*. Amsterdam: Elsevier Science Ltd., pp. 291-306.

Balzer, Wolfgang, Moulines, C. Ulises, Sneed, Joseph D., (1987): *An Architectonic for Science*. Dordrecht: D. Riedel Publishing Company.

Balzer, W, Moulines, C.U., and Sneed, J.D (Eds.) (2000): *Structuralist Knowledge Representation: Paradigmatic Examples*. *Poznan Studies in the Philosophy of the Sciences and the Humanities Vol. 75.* Amsterdam: Rodopi.

Barcan, Marcus et al., (Eds.) (1986): *Proceedings of the 7th International Congress of Logic, Methodology, and Philosophy of Science*. Amsterdam: Elsevier Science Ltd.

Bechtel, William (1988): *Philosophy of Mind: An Overview for Cognitive Science*. New Jersey: Lawrence Erlbaum Associates.Inc.

Bechtel, William and Abrahamsen, Adele (2002): *Connectionism and the Mind 2nd Ed*. Hongkong: Blackwell Publishers.

Bechtel William and Hamilton, Andrew (2007): Reduction, Integration, and the Unity of Science: Natural, Behavioral, and Social Sciences and the Humanities in Kuipers (ed.) (2007): *Philosophy of Science: Focal Issues (The Handbook of the Philosophy of Science Vol. 1*. New York: Elsevier Science Ltd, pp. 377-430.

Berkowitz L. (1984): *Advances in Experimental Social Psychology Vol 17*. NewYork: Academic, p. 229-266.

Bickle, John (1993): "Connectionism, Eliminativism, and the Semantic View of Theories." In *Erkenntnis* 39, 1993, no. 5, Pp. 359-382.

Bickle, John (1998): *Psychoneural Reduction: The New Wave*. Cambridge, Massachusetts: A Bradford Book, MIT Press.

Birnbaum, P. H. (1977): Assessment of alternative management forms in interdisciplinary projects. In *Management Science* 24 (3), p. 272–84.

Borgelt, C. et. al. (2003): *Neuro-Fuzzy-Systeme: Von den Grundlagen künstlicher Neuronaler Netze zur Kopplung mit Fuzzy-Systemen 3. Auflage*. Wiesbaden: Vieweg.

Bourbaki, N. (2004): *Element of Mathematics: Theory of Set*. Berlin: Springer.

Bungartz, Hans-Joachim, Zimmer, Stefan, Buchholz, Martin and Pflüger, Dirk (2014): *Modeling and Simulation: An Application-oriented Introduction*. Heidelberg: Springers.

Carnap, R. (1928): *Der logische Aufbau der Welt*. Hamburg: Felix Meiner, 1999.

Carnap, R. (1934): *The Unity of Science*, Oxon: Routledge, 2013.

Cart, Jordi (2013): The Unity of Science. In Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/scientific-unity/#Red.

Churchland, Patricia S., and Sejnowski, Terrence J. (1989): Neural Representation and Neural Computation. In Lycan, William G. and Prinz, Jesse J. (Eds.) (2011): *Mind and Cognition: An Anthology 3rd Ed*. Malden: Blackwell Publishing, pp. 247-268.

Churchland, Patricia S., Sejnovki, Terence J. (1996): *The Computational Brain*. Cambridge, Massachusetts: A Bradford Book, MIT Press.

Churchland, Paul M. (1991): Folk psychology and explanation of Human Behavior. in Greenwood, John D. (2008): *The Future of Folk Psychology: Intentionality and Cognitive Science* revised Edition. Canada: Cambridge University Press, p. 51-69.

Churchland, Paul M. (1997): *Die Seelen Maschine*. Heidelberg: Spektrum.

Cooper, J. & Fazio, R.H. (1984): A New Look at Dissonance Theory. In Berkowitz L. (1984): *Advances in Experimental Social Psychology Vol 17*. NewYork: Academic, p. 229-266.

Dennett, Daniel C. (1981): True Believers: The Intentional Strategy and Why it Works. In Lycan, William G. and Prinz, Jesse J. (Eds.) (2011): *Mind and Cognition: An Anthology 3rd Ed*. Malden: Blackwell Publishing, pp. 323-336.

Dizadji-Bahmani, F., Frigg, R., and Hartmann S. (2009): Who's Afraid of Nagelian Reduction? In *Erkenntnis* 73, 2010, pp. 393-412.

Dizadji-Bahmani, F., Frigg, R., and Hartmann S. (2011): Confirmation and Reduction: a Bayesian Account. In *Synthese* 179, 2011, pp. 321-338.

Dupré, John (1995): *The Disorder of Things*. Cambridge, Massachusetts: Harvard University Press.

Echeverria, Javier, Ibarra, Andoni and Mormann, Thomas (ed.) (1992): *The Space of Mathematics: Philosophical, Epistemological, and Historical Explorations*. Berlin: de Gruyter, pp. 403-411.

Egan, M. Frances (1989): What's Wrong with the Syntactic Theory of Mind. In *Philosophy of Science* 56 (4), 1989, pp. 664-674.

Ezquerro, Jesús, and Manrique, Fernando, M. (2004): Intertheory Relations in Cognitive Science: Priviledged Levels and Reductive Strategies. In *Crítica: Revista Hispanoamericana de Filosofía* 36 (106), 2004, pp. 55-103.

Festinger, Leon (1985): *A Theory of Cognitive Dissonance*. Stanford: Stanford University Press.

Forge, John (2002): *Reflections on Structuralism and Scientific Explanation.* in *Synthese* 130, 2002, pp. 109-121.

Fornito, Alex, Zalesky, Andrew, and Bullmore, Edward (2016): *Fundamentals of Brain Network Analysis*. London: Academic Press, Elsevier Science Ltd.

Freedman, J.L. (1965): Long-term Behavioral Effects of Cognitive Dissonance. In *Journal of Experimental Social Psychology* 1(2), 1965, pp. 145-155

Friedenberg, Jay, and Silverman, Gordon (2012): *Cognitive Science: An Introduction to the Study of Mind, 2nd Ed.* California: SAGE Publication, Inc.

Frodeman, Robert., Klein, Julie Thompson, and Mitcham, Carl.(2010): *The Oxford Handbook of Interdisciplinarity*. New York: Oxford University Press.

Funtowicz, S.O. and Ravetz, J.R. (1993). The emergence of post-normal science. In R. von Schomberg (ed.) (1993): *Science, Politics, and Morality: Scientific Uncertainty and Decision Making*. Dordrecht: Kluwer Academic Publishers. pp. 85–123.

Gazzaniga, Michael S., Ivry, Richard B., Mangun, George R. (1998): *Cognitive Neuroscience: The Biology of the Mind.* New York: W.W. Norton & Company.

Gähde, Ulrich (1996): Holism and the Empirical Claim of Theory-Nets. In Balzer, Wolfgang, et al. (eds.): *Structuralist Theory of Science: Focal Issues, New Results*, in *Perspectives in Analytical Philosophy*, Bd. 6. Berlin: de Gruyter, pp. 167-190.

Gähde, Ulrich: Modelle der Struktur und Dynamik wissenschaftlicher Theorien. In Bartels, Andreas, Stöckler, Manfred (Eds.) (2007): *Wissenschaftstheorie. Ein Studienbuch*. Paderborn: Mentis, pp. 45-65.

Gibbons, M. et al. (1994): *The new production of knowledge*. London: Sage.

Goldstein, E. Bruce (2010): *Sensation and Perception*. Belmont, CA: Wadsworth Cengage Learning.

Gordon, Robert M. (1986): Folk Psychology as Simulation. In Lycan, William G., and Prinz, Jesse J. (eds.) (2011): *Mind and Cognition: An Anthology*. Oxford: Blackwell Publishing. pp.369-378.

Greenwood, John D. (2008): *The Future of Folk Psychology: Intentionality and Cognitive Science* revised Edition. Canada: Cambridge University Press, pp. 51-69.

Haare, Rom (2002): *Cognitive Science: a Philosophical Introduction*. London: Sage Publication. Ltd.

Harmon-Jones, Eddie, and Harmon-Jones (2007): Cognitive Dissonance Theory After 50 Years of Development. In *Zeitschrift für Social Psychologie* 38 (1), 2007, pp. 6-17.

Harmon-Jones, Eddie & Mills, Judson (eds.) (2009): *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology.* Washington D.C.: American Psychological Association.

Haykin, Simon (2009): *Neural Networks and Learning Machine, 3 Ed*. New Jersey: Pearson Education.

Hawkins, Robert D. and Kandel, Eric R. (1984a): Is There a Cell-Biological Alphabet for Simple Forms of Learning? In *Psychological Review* 91 (3), 1984, pp. 375-391.

Hawkins, Robert. D., and Kandel, Eric. R. (1984b): Steps Toward a Cell-Biological Alphabet for Elementary Forms of Learning. In G. Lynch, J. L. McGaugh, and N. M. Weinberger, (eds.) (1984): *Neurobiology of Learning and Memory*. New York: Guilford Press, pp. 385-404.

Heckhausen, Heinz (1987): 'Interdisziplinäre Forschung' zwischen Intra-, Multi- und Chimären-Disziplinarität. In Kocka, Jürgen (ed.)(1987): *Interdisziplinarität. Praxis – Herausforderung – Ideologie*. Frankfürt am Mainz: Suhrkamp, pp. 129-145.

Hohwy, Jakob and Kallestrup, Jesper (2008): *Being Reduced*. New York: Oxford University Press Inc, pp. 93-114.

Hooker, C. A. (2004): Asymptotics, Reduction and Emergence in *The Britisch Journal for the Philosophy of Science* 55 (3), 2004, pp. 435-479.

Hopfield, John. J. (1982): Neural Networks and Physical Systems with Emergent Collective Computational Abilities. In *Proceedings of the National Academy of Sciences* (79) 1982, pp. 2554-2558.

Horgan, Terence, and Woodward, James (1996): Folk Psychology is Here to Stay. In Lycan, William G. and Prinz, Jesse J. (Eds.) (2011): *Mind and Cognition: An Anthology 3rd Ed*. Malden: Blackwell Publishing, pp.419-436.

Jarcho, Johanna M., Berkman, Elliot T., and Lieberman, Matthew D. (2010): The Neural Basis of Rationalization: Cognitive Dissonance Reduction During Decision-Making. In *SCAN* 6, 201I, pp. 460-467.

Jungert, Michael., Romfeld, Elsa., Sukopp, Thomas., Voigt, Uwe (Eds.) (2013): *Interdisciplinarität: Theorie, Praxis, Probleme, 2. Auflage*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Kamlah, Andreas (1985): On Reduction of Theories. In *Erkenntnis* 22 (1), 1985, pp. 119-142.

Kandel, Eric R., Schwarz, James H., Jessell, Thomas M. Eds. (1995): *Neurowissenschaften: Eine Einführung*. Heidelberg: Spektrum.

Kandel, Eric R., Schwarz, James H., Jessell, Thomas M., Siegelbaum, Steven A., and Hudspeth, A.j. Eds. (2013): *Principles of Neural Science 5th Ed*. New York: McGraw-Hill Medical.

Kim, Jaegwon (2008): Reduction and Reductive Explanations: Is One Possible Without the Other? In Hohwy, Jakob and Kallestrup, Jesper (2008): *Being Reduced*. New York: Oxford University Press Inc, pp. 93-114.

Kleene, S.C. (1951): *Representation of Events in Nerve Nets and Finite Automata* in http://www.dlsi.ua.es/~mlf/nnafmc/papers/kleene56representation.pdf.

Klein, Julie Thompson. (1990): *Interdisciplinarity: History, Theory, & Practice*. Detroit:Wayne State University Press.

Klein, Julie Thompson (2010): A Taxonomy of Interdisciplinarity. In Frodeman, Robert., Klein, Julie Thompson, and Mitcham, Carl. (eds.) (2010): *The Oxford Handbook of Interdisciplinarity*. New York: Oxford University Press, pp. 18-30.

Koch, Christof, and Segev, Idan (2001): *Methods in Neuronal Modeling: From Ions to Networks 2nd edition*. Cambridge, Massachusetts: MIT Press.

Kocklemans, Joseph J. (ed.) (1979): *Interdisciplinarity and Higher Education*. University Park, Pennsylvania: Pennsylvania State University Press.

Kuipers (ed.) (2007): *Philosophy of Science: Focal Issues (The Handbook of the Philosophy of Science Vol. 1*. New York: Elsevier Science Ltd.

Lakatos, Imre (1978): *The Methodology of Scientific Research Programmes: Philosophical Papers Vol 1*. Cambridge: Cambridge University Press, 1980.

Lattuca, Lisa R. (2001): *Creating Interdisciplinarity: Interdisciplinary Research and Teaching among College and University Faculty.* Nashville: Vanderbilt University Press.

Lauth, Bernhard (Stand Feb 2015): *Neurokognition: Skript zum HS Neuronale Netze im WS 2014/15*.

Link, Godehard (2009): *Collegium Logicum: Logische Grundlage der Philosophie und der Wissenschaften band 1*. Paderborn: Mentis.

Luger, George F. (1994): *Cognitive Science: The Science of Intelligent Systems*. San Diego: Academic Press.

Lycan, William G., and Prinz, Jesse J. eds (2011): *Mind and Cognition: An Anthology*. Oxford: Blackwell Publishing.

Lynch, Gary, McGaugh, James L., and Weinberger, Norman M. (1984): *Neurobiology of Learning and Memory.* New York: Guilford Press.

McCulloch, Warren S., and Pitts, Walter (1943): A Logical Calculus of the Ideas Immanent in Nervous Activity. In *Bulletin of Mathematical Biophysics* 5, 1943, pp. 115-133.

McKay, Thomas, and Nelson, Michael (2010): The de dicto/de re Distinction. In *Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/prop-attitude-reports/dere.html.

McLeod, Peter, et. al. (1998): *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.

McLeod, Saul (2014): Cognitive Dissonance. In *SimplyPsychology* https://www.simplypsychology.org/cognitive-dissonance.html.

Mendelson, Bert (1990): *Introduction to Topology 3 ed*. New York: Dover Publications.

Mitchell, Sandra D. (2003): *Biological Complexity and Integrative Pluralism*. New York: Cambridge University Press.

Mitchell, Sandra D. (2009): *Unsimple Truths*. Chicago: The University of Chicago Press.

Morrison. Margaret (2006): Emergence, Reduction, and Theoretical Principles: Rethinking Fundamentalism. In *Philosophy of Science* 73, 2006, pp. 876-887.

Moulines, C. Ulises (1984a): Ontological Reduction in the Natural Science (1) In Balzer W. et al. (eds.) (1984): *Reduction in Science*. Dordrecht: D. Riedel Publishing Company, pp. 51-70.

Moulines, C. Ulises (1984b): *Links, Loops, and the Global Structure of Science* in *Philosophia Naturalis* 21, 1984, pp. 254-265.

Moulines, C. Ulises (1992): *Towards a Typology of Intertheoretical Relations* in Echeverria, Javier, Ibarra, Andoni, and Mormann,

Thomas (eds.) (1992): *The Space of Mathematics: Philosophical, Epistemological, and Historical Explorations*. Berlin: de Gruyter, pp. 403-411.

Moulines, C. Ulises (1996): *Structuralism: The Basic Ideas* in Balzer, Wolfgang, and Moulines, C.U. (eds.) (1996): *Structuralist Theory of Science: Focal Issues, New Results*, in *Perspectives in Analytical Philosophy*, Bd. 6. Berlin: de Gruyter, pp. 1-14.

Moulines, C. Ulises (1997): The Concept of Universe from a Meta-theoretical Point of View. In Ibarra, Andoni and Mormann, Thomas (1997): *Representations of Scientific Rationality: Poznan Studies in Philosophy of the Sciences and Humanities Vol. 6*. Amsterdam: Rodopi, pp. 359-379.

Moulines, C. Ulises (2008): *Die Entwicklung der modernen Wissenschaftstheorie (1890-2000)*. Hamburg: LIT Verlag.

Moulines, C.Ulises (2014): Intertheoretical Relations and the Dynamics of Science. In *Erkenntnis* 79, 2014, pp. 1505-1519.

Moulines, C. Ulises, and Polanski, Marek. (1996): *Bridges, Constraints, and Links* in Balzer, Wolfgang, et al. (Eds.) (1996): *Structuralist Theory of Science: Focal Issues, New Results*, in *Perspectives in Analytical Philosophy*, Bd. 6. Berlin: de Gruyter, pp. 219-233.

Nagel, Ernst (1979): *The Structure of Science. Problems in the Logic of Explanation*, Indianapolis: Hackett Publishing Company, Inc.

Organisation for Economic Cooperation and Development (OECD) (1972): *Interdisciplinarity: Problems of Teaching and Research in Universities*. Paris: Author.

Palmer, Carole L (1996): Introduction. In *Library Trends* 45 (2), 1996, pp. 129-33.

Piccinini, Gualtiero (2004): The First Computational Theory of Mind and Brain: A Close Look at McCulloch and Pitt's "Logical Calculus of Ideas Immanent in Nervous Activity" in *Synthese,* 141, 2004, pp. 175-215.

Ravenscroft, Ian (2010): Folk Psychology as a Theory. In *Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/folkpsych-theory/.

Rosenberg, M.J. and Hovland, C.I. (1960): Cognitive, Affective and Behavioral Components of Attitudes. In Rosenberg, M.J. and Hovland, C.I., Eds. (1960): *Attitude Organization and Change: An Analysis of Consistency among Attitude Components*. New Haven: Yale University Press, pp. 1-14.

Schaffner, Kenneth F. (1967): Approaches to Reduction. In *the Philosophy of Science Association* 34 (2), 1967, pp. 137-147.

Seung, Sebastian (2012): *Connectome: How the Brain's Wiring Makes Us Who We Are*. Houghton Mifflin Harcourt. Kindle-Version.

Shultz, Thomas R (2003): *Computational Developmental Psychology.* Cambridge: MIT Press.

Shultz, Thomas R., and Lepper, Mark R. (1992): A Constraint Satisfaction Model of Cognitive Dissonance Phenomena. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society. New York: Psychology Press*, pp. 462-467.

Shultz, Thomas R., and Lepper, Mark R. (1996): Cognitive Dissonance Reduction as Constraint Satisfaction. In *Psychological Review* 103 (2), 1996. pp. 219-240.

Shultz, Thomas R., and Lepper, Mark R. (1998): The Consonance Model of Dissonance Reduction. In Read, Stephen John and Miller Lynn C.

(eds.) (2013): *Connectionist Models of Social Reasoning and Social Behaviour.* Hove: Routledge, pp. 211-244.

Shultz, Thomas R., and Lepper, Mark R. (1999): Computer Simulation of Cognitive Dissonance Reduction. In Harmon-Jones, Eddie and Mills, Judson (eds.) (2009): *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology.* Washington D.C.: American Psychological Association, pp. 235-265.

Shultz, Thomas R., and Lepper, Mark R. (2000): Simulation of Self-affirmation Phenomena in Cognitive Dissonance. https://escholarship.org/uc/item/78t7t7t6.

Sklar, Lawrence (1967): Types of Inter-Theoretic Reduction. In *the British Journal for the Philosophy of Science* 18 (2), 1967, pp. 109-124.

Sporns, Olaf (2012): *Discovering the Human Connectome*. Cambridge, Massachusetts: MIT Press.

Sporns, Olaf (2016): *Networks of the Brain*. New York: MIT Press.

Stephan, Achim and Walter, Sven (2013): *Handbuch Kognitions-wissenschaft*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Stich, Stephen P. (1978): Autonomous Psychology and the Belief-Desire Thesis. In Lycan, William G. and Prinz, Jesse J. (eds.) (2011): *Mind and Cognition: An Anthology 3rd Ed*. Malden: Blackwell Publishing, pp. 405-418.

Stich, Stephen P. (1981): Dennett on Intentional Systems. In Lycan, William G. and Prinz, Jesse J. (eds.) (2011): *Mind and Cognition: An Anthology 3rd Ed.* Malden: Blackwell Publishing, pp. 337-350.

Stich, Stephen (1991): *From Folk Psychology to Cognitive Science: The Case against belief. Cambridge:* MIT Press.

Stich, Stephen P., and Ravenscroft, Ian (1993): What is Folk Psychology? In *Technical Report #5*, 1993. New Jersey: Rutgers University Center for Cognitive Science.

Stichweh, R. (1984): *Zur Entstehung des modernen Systems wissenschaftlicher Disziplinen: Physik in Deutschland 1740–1890*. Frankfurt am Main: Suhrkamp.

Sukopp, Thomas (2013): Interdisciplinarität und transdisciplinarität. In Jungert, Michael., Romfeld, Elsa., Sukopp, Thomas., Voigt, Uwe (eds.) (2013): *Interdisciplinarität: Theorie, Praxis, Probleme, 2. Auflage*. Darmstadt: Wissenschaftliche Buchgesellschaft., Pp. 13-29.

Suppes, Patrick (1999): *Introduction to Logic.* Mineola, NY: Dover Publishings.

Thagard, Paul (2005): *Mind: Introduction to Cognitive Science*. Cambridge, MA: MIT Press.

Thagard, Paul (2014): Cognitive Science. In *Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/cognitive-science/.

van Leeuwen, Theo. (2005): Three models of interdisciplinarity. In Wodak, Ruth, & Chilton, Paul (eds.) (2005): *A New Agenda in (Critical) Discourse Analysis: Discourse Approaches to Politics, Society and Culture*. Amsterdam: John Benyamins Publishing Co., pp. 3-18.

Van Overwalle, Frank (2015): *Social Connectionism: A Reader and Handbook for Simulations*. New York: Psychology Press.

Van Overwalle, Frank, and Jordens, Karen (2002): An Adaptive Connectionist Model of Cognitive Dissonance. in *Personality and Social Psychology Review* 6 (3), 2002, pp. 204-231.

Van Riel, Raphael (2011): Nagelian Reduction beyond the Nagel Model. In *Philosophy of Science* 78 (3), 2011, pp. 353-375.

Van Riel, Raphael and van Gulick, Robert (2014): Scientific Reduction. In Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/scientific-reduction/.

Van Veen, Vincent, and Carter, Cameron S. (2006): Conflict and Cognitive Control in the Brain. In *Association for Psychological Science* 15 (5), 2006, pp. 237-240

Van Veen, Vincent, Krug, Marie K., Schooler, Jonathan W., and Carter, Cameron S. (2009): Neural Activity Predicts Attitude Change in Cognitive Dissonance. In *Nature Neuroscience* 12 (11), 2009, pp. 1469-1474.

Voigt, Uwe (2013): Interdisciplinarität: Ein Model der Modelle. In Jungert, Michael., Romfeld, Elsa., Sukopp, Thomas., Voigt, Uwe (eds.) (2013): *Interdisciplinarität: Theorie, Praxis, Probleme, 2. Auflage*. Darmstadt: Wissenschaftliche Buchgesellschaft, pp. 31-46.

Weingart, Peter. (2010): A Short History of Knowledge Formations. In Frodeman, Robert., Klein, Julie Thompson, and Mitcham, Carl. (eds.) (2010): *The Oxford Handbook of Interdisciplinarity*. New York: Oxford University Press, pp. 3-14.

Westermann, Rainer (1989): Festinger's Theory of Cognitive Dissonance: A Revised Structural Reconstruction. In Westermeyer, Hans (ed.) (1989): *Psychological Theories from a Structuralist Point of View*. Berlin: Springer, pp. 33-62.

Westermann, Rainer (2000): Festinger's Theory of Cognitive Dissonance: A Structuralist Theory-Net. In Balzer, W, Moulines, C.U., and Sneed, J.D (2000): *Structuralist Knowledge Representation: Paradigmatic Examples. Poznan Studies in the Philosophy of the Sciences and the Humanities Vol. 75*. Amsterdam: Rodopi, pp. 189-217.

Westermeyer, Hans (ed.) (1989): *Psychological Theories from a Structuralist Point of View*. Berlin: Springer.

Wilson, Robert A., and Keil, Frank C. (eds.) (1999): *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, Massachusetts: MIT Press.

Wodak, Ruth., & Chilton, Paul. (eds.) (2005): *A New Agenda in (Critical) Discourse Analysis: Discourse Approaches to Politics, Society and Culture*. Amsterdam: John Benyamins Publishing Co.

Zenker, Frank, Gärdenfors, Peter (2012): Modeling Diachronic Changes in Structuralism and in Conceptual Spaces.