Reasoning with False Evidence

Zur Erlangung des akademischen Grades einer
DOKTORIN DER PHILOSOPHIE (Dr. phil)

von der KIT Fakultät für Geistes- und Sozialwissenschaften des
Karlsruher Instituts für Technologie (KIT)
angenommene

DISSERTATION

von

Corinna Günth-Stiegeler

KIT-Dekan: Prof. Dr. Michael Mäs

1. Gutachter: Prof. Dr. Gregor Betz
2. Gutachter: Prof. Dr. Dr. Claus Beisbart

Tag der mündlichen Prüfung: 08.03.2023

Contents

Zusammenfassung

Unter heutigen Wissenschaftsphilosophen sind die folgenden zwei Annahmen recht unkontrovers. Erstens, empirische Evidenzen sind fallibel und viele unserer heutigen und gestrigen empirischen Überzeugungen sind bzw. waren falsch. Zweitens, unsere Fähigkeit korrekt auf die Wahrheit oder Falschheit einer Hypothese zu schließen hängt ab von der Korrektheit unserer Evidenzmenge. Ausgehend von diesen zwei Annahmen stellt sich die Frage, wie Wissenschaftler mit falschen Evidenzen vernünftig argumentieren und zuverlässig auf Hypothesen schließen können.

In wissenschaftlichen Debatten werden Überzeugungen oft mittels des Konzepts der Relevanz-Bestätigung gebildet. Dieses Konzept gehört in den Bereich der Bayesschen Bestätigungstheorie, welche eine probabilistische Bestätigungstheorie ist. Das Maß der Bestätigung, das eine Hypothese durch eine Evidenz erfährt, ist im Falle der Relevanz-Bestätigung eine Funktion der absoluten Wahrscheinlichkeit der Hypothese sowie der bedingten Wahrscheinlichkeit der Hypothese gegeben die Evidenz. Eine andere Form der Bestätigung ist die Absolut-Bestätigung. In diesem Fall ist das Bestätigungsmaß lediglich eine Funktion der bedingten Wahrscheinlichkeit der Hypothese gegeben die Evidenz. Es existiert eine beeindruckende Vielfalt an diskutierten Bayesschen Bestätigungsmaßen. Aber nur drei von ihnen lassen sich aus guten Gründen als eine Erweiterung des Konzepts des deduktiven Schließens verstehen. Diese guten Gründe finden sich meines Erachtens in (Crupi and Tentori, 2013) bzw. (Crupi et al., 2007). Diese drei Bestätigungsmaße sind die bedingte Wahrscheinlichkeit der Hypothese gegeben die Evidenz, $P(h|e)$, sowie $Z_P(h, e)$ und $F_P(h, e)$. Die beiden letztgenannten sind Relevanz-Bestätigungsmaße und wurden zuerst definiert von Crupi and Tentori (2010) und Branden Fitelson (2004) bzw. Kemeny and Oppenheim (1952). Das erstgenannte ist ein Absolut-Bestätigungsmaß. In dieser Arbeit wird jede Analyse für alle drei Bestätigungsmaße durchgeführt.

Diese Arbeit verwendet einen dialektischen Wahrscheinlichkeitsbegriff. Wie in (Betz, 2010) gezeigt, erfüllt das Maß an Begründung, das eine Hypothese durch eine Evi-

denzmenge gegeben eine dialektische Struktur erfährt, $DOJ(h|e)$, die Kolmogorovschen Axiome und ist somit eine Wahrscheinlichkeit. Eine dialektische Struktur besteht aus Sätzen und deduktiv gültigen Argumenten, die sich gegebenenfalls gegenseitig stützen oder angreifen. Gemäß der Theorie dialektischer Strukturen kann jedes Stadium einer Debatte dargestellt werden durch eine dialektische Struktur und die Positionen ihrer Teilnehmer. Die Theorie dialektischer Strukturen ist ein formales Modell komplexer Argumentation, das ausführlich vorgestellt wird in (Betz, 2010).

Die eigenen Überzeugungen mittels des Konzepts der Bestätigung zu bilden, ist das eine wertvolle Art und Weise zu Schlussfolgern, falls manche meiner Evidenzen falsch sind? Welche verschiedenen Arten, die eigenen Überzeugungen anhand des Konzepts der Bestätigung zu bilden, gibt es? Sind sie unterschiedlich wertvoll, falls der Wert darin besteht nur wahre Aussagen zu akzeptieren? Wie wertvoll in diesem Sinne ist es, den Grad der eigenen Überzeugungen dem Begründungsgrad anzupassen? Wie wertvoll in diesem Sinne ist es, nur diejenigen Hypothesen zu akzeptieren, die ausreichend durch die eigene Evidenz bestätigt sind? Diese Fragen lösen eine Hinwendung zur Statistik aus und werden detailliert im ersten Teil meiner Arbeit beantwortet, das heißt in Kapitel 2.

Was passiert in tatsächlichen wissenschaftlichen Debatten? Wie ändern Teilnehmer ihre Überzeugungen so, dass sie letztlich einen Konsens erreichen? Wechseln sie von einem Paar von Hypothese und Evidenzmenge zu einem anderen derart, dass der Begründungsgrad erhöht wird? Wechseln sie von einem Paar von Hypothese und Evidenzmenge zu einem anderen derart, dass es keine andere Hypothese gibt, die in einem höheren Maße bestätigt wird? Erhöhen Wechsel dieser Art immer die Ähnlichkeit mit dem finalen Konsens? Diese Fragen führen zu einer Hinwendung zur Wissenschaftsgeschichte und werden detailliert im zweiten Teil meiner Arbeit beantwortet, das heißt in Kapitel 3.

In beiden Teilen meiner Arbeit benötigt die Analyse von Debatten hinsichtlich des Bestätigungsgrades von Hypothesen die Hilfe von Computern, sowohl bestimmte computationale Techniken als auch Leistungen. Damit trägt diese Arbeit zu einem relativ neuen Philosophiezweig bei, nämlich dem der computerunterstützten Philosophie. Gemäß Grim and Singer (2020) beinhaltet dieser Zweig all jene philosophischen Unternehmungen, die von Computern Gebrauch machen. In beiden Teilen meiner Arbeit werden Debatten als dialektische Strukturen dargestellt und hinsichtlich der Bestätigung von Hypothesen analysiert. Das Konzept der Bestätigung wird mittels

des Konzepts der Begründung ausbuchstabiert. Dieses ist wiederum der Theorie dialektischer Strukturen entnommen. Ein nicht zu unterschätzender Vorzug dieser Theorie ist ihre Anbindefähigkeit an computationale Analysen. Diese liegt darin begründet, dass es sich bei einer dialektischen Struktur um eine Boolesche Funktion handelt.

Der erste Teil meiner Arbeit nutzt 1000 simulierte Debatten von Betz (2013) sowie Monte-Carlo-Methoden. Im zweiten Teil meiner Arbeit nutze ich eine Argumentationssoftware, nämlich Argdown von Christian Voigt (2018), um dialektische Strukturen zu implementieren, die den verschiedenen Phasen einer historischen Debatte entsprechen. In beiden Fällen werden computationale Techniken und Ressourcen benötigt um Begründungsgrade zu berechnen. Für alle dialektischen Strukturen dieser Arbeit gilt, dass die Berechnung des Begründungsgrades einer Hypothese menschliche Fähigkeiten bei Weitem übersteigt. Alle meine computationalen Analysen werden mit Hilfe des Computeralgebraprogramms Mathematica von Wolfram Research, Inc. (2019) durchgeführt, unter Verwendung der Ressourcen des Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT) (2017) und KIT's Department of Philosophy, ITZ and ITAS (2017). Eine Sammlung meines Analysecodes sowie alle dialektischen Strukturen der von mir rekonstruierten historischen Debatte finden sich unter https://github.com/cguenth/CONFasVI.git bzw. https://github.com/cguenth/GDC.git.

In beiden Teilen meiner Arbeit zeigen sich Unterschiede zwischen Absolut- und Relevanz-Bestätigung. Und ebenso zeigt sich in beiden Teilen meiner Arbeit, dass Regeln niemals ausnahmslos gelten. Dies bestätigt Feyerabend (1976), welcher bestreitet, dass es ausnahmslos gültige wissenschaftliche Regeln gibt. Jedoch zeigt meine Arbeit nicht, dass es überhaupt keine Regeln gibt. Ausnahmen lassen sich in beiden Teilen meiner Arbeit mittels des Konzepts der Evidenz höherer Ordnung charakterisieren.

Was ist Evidenz höherer Ordnung? Zunächst möchte ich meine eigene Auffassung dieses Konzepts vorstellen. In dieser Arbeit wird eine Evidenzmenge als Evidenz erster Ordnung betrachtet und Aussagen über eine Evidenz erster Ordnung als Evidenz höherer Ordnung. Beispiele hierfür sind Aussagen über die Größe und Korrektheit der Evidenzmenge, sowie die Eigenschaften der argumentativen Struktur, in welche die Evidenzmenge eingebettet ist. Die Analysen dieser Arbeit beziehen sich auf drei bestimmte Beispiele, nämlich die inferentielle Dichte der dialektischen Struktur

zu einem gewissen Zeitpunkt und, für jeden Debattenteilnehmer, die Menge aller Evidenzen sowie der Anteil wahrer Evidenzen.

Die Natur von Evidenz höherer Ordnung wird heutzutage lebhaft diskutiert. Es gibt wenigstens zwei unterschiedliche Auffassungen. Zum Einen wird Evidenz höherer Ordnung aufgefasst als Evidenz über den Charakter der Evidenz. Zum Anderen wird Evidenz höherer Ordnung verstanden als die Fähigkeit eines Subjekts der Evidenz gegenüber vernünftig zu reagieren. Gemäß Talbott (2016b) ist Evidenz höherer Ordnung Evidenz besonderer Art, da sie die Zuverlässigkeit eines Schlusses untergräbt. Beispiele dafür, wie die Zuverlässigkeit eines Schlusses untergraben werden kann, finden sich in (Pollock, 1984).

Die Auffassung von Talbott (2016b) lässt sich mit meiner eigenen Auffassung von Evidenz höherer Ordnung verknüpfen. Zu diesem Zweck sei angenommen, dass die Teilnehmer einer Debatte ihre Überzeugungen nach dem Begründungsgrad ausrichten. In diesem Fall ist Evidenz höherer Ordnung im Sinne von Talbott (2016b) all jenes, welches die Zuverlässigkeit dieses kognitiven Prozesses untergräbt. Betz (2015) hat gezeigt, dass die Zuverlässigkeit von Bestätigung als veritistischer Indikator von Evidenz höherer Ordnung in meinem Sinne abhängt, nämlich der inferentiellen Dichte und der Größe der Evidenzmenge. Diese Ergebnisse sind jedoch limitiert. Erstens geht Betz (2015) von einer vollständig korrekten Evidenzmenge aus. Zweitens betrachtet Betz (2015) nur Absolut-Bestätigung als veritistischen Indikator. Der erste Teil meiner Arbeit erweitert (Betz, 2015) nicht nur durch die Berücksichtigung nur teilweise korrekter Evidenzmengen, sondern auch durch die Berücksichtigung des Konzepts der Relevanz-Bestätigung.

In dieser Arbeit wird die Zuverlässigkeit eines veritistischen Indikators bemessen anhand eines statistischen Hypothesentests auf der Basis Monte-Carlo-generierter Daten. Die Zuverlässigkeit eines veritistischen Indikators erhöht sich, falls die Signifikanz abnimmt oder die Mächtigkeit des Tests zunimmt. Die Hypothese $\neg h$ wird gegen die Alternative h getestet. Für jeden statistischen Test existiert eine kritische Region, das heißt eine Region, in der die Hypothese $\neg h$ verworfen werden muss. Diese kritische Region wird bestimmt, indem die Mächtigkeit auf 0.25 festgelegt wird. Diese Arbeit erweitert (Betz, 2015) also auch durch eine neue Art und Weise der Bestimmung der Zuverlässigkeit von Bestätigung als veritistischer Indikator.

Unter heutigen Wissenschaftsphilosophen werden einige sogenannte Prinzipien der Überzeugungsrevision diskutiert. Diese empfehlen eine Neubewertung der eigenen

Überzeugungen im Lichte bestimmter Evidenzen hinsichtlich der Zuverlässigkeit des Prozesses, welcher den Überzeugungen zugrunde liegt. Der erste Teil meiner Arbeit untermauert gewisse Prinzipien der Überzeugungsrevision, wie zum Beispiel das $MERF$-Prinzip von Talbott (2016b) oder das Integrations-Prinzip von Christensen (2008), auf zweifache Weise. Erstens bestimmt es die Zuverlässigkeit spezieller kognitiver Prozesse, nämlich bestimmter Formen der Überzeugungsbildung anhand des Konzepts der Bestätigung. Zweitens identifiziert es hinsichtlich der Zuverlässigkeit relevante Kategorien dieses Prozesses unter Verwendung des Konzepts der Evidenz höherer Ordnung. Das Integrations-Prinzip verlangt, dass die Überzeugungen auf Objektebene die Überzeugungen auf Metaebene bezüglich der Zuverlässigkeit des kognitiven Prozesses widerspiegeln. Das $MERF-$Prinzip fordert eine Überzeugungsrevision, falls eine Überzeugung nicht der Zuverlässigkeit des kognitiven Prozesses entspricht (und es keine Kategorisierung dieses Prozesses gibt, für die das doch zumindest annähernd der Fall ist). Talbott (2016b) bemisst die Zuverlässigkeit eines kognitiven Prozesses anhand der relativen Häufigkeit wahrer Überzeugungen innerhalb der Menge aller Überzeugungen, die von diesem Prozess generiert werden.

Nicht nur die Natur von Evidenz höherer Ordnung wird heutzutage lebhaft diskutiert, sondern auch die Tragweite dieses Konzepts. Gibt es zum Beispiel eine Situation, in welcher Evidenz von Evidenz (für eine Hypothese h) selbst Evidenz für h ist? Von manchen Epistemologen wird diese Frage bejaht, vergleiche zum Beispiel (Feldman, 2005). Manch andere Epistemologen bezweifeln zumindest die bisher angeführten Beispiele für solche Situationen, siehe zum Beispiel (Fitelson, 2012). Es sei angenommen, dass (i) das Maß der Bestätigung ein Beispiel für eine Evidenz höherer Ordnung ist und (ii) dass es eine Situation gibt, in welcher Bestätigung ein zuverlässiger veritistischer Indikator ist. Dann gibt es eine Situation, in welcher Evidenz von Evidenz (für eine Hypothese h) selbst Evidenz für h ist.

Die wichtigsten Resultate des ersten Teils meiner Arbeit sind:

- ($V1$) Die relative Häufigkeit wahrer Hypothesen innerhalb der Menge aller Hypothesen mit einem bestimmten Bestätigungsgrad. Es zeigt sich, dass, für eine Evidenzmenge ohne falsche Aussagen, die relative Häufigkeit nur dann von Evidenz höherer Ordnung abhängt, falls das Konzept der Relevanz-Bestätigung verwendet wird.
- ($V2$) Die Zuverlässigkeit von Bestätigung als veritistischer Indikator. Unabhängig von Evidenz höherer Ordnung zeigt sich, dass Absolut-Bestätigung ein

zuverlässigerer veritistischer Indikator ist als Relevanz-Bestätigung. Die Abhängigkeit der Zuverlässigkeit von Evidenz höherer Ordnung ist nicht immer dieselbe für Absolut- und Relevanz-Bestätigung:

  - ($V2.0$) Mit zunehmendem Anteil falscher Aussagen in der Evidenzmenge nimmt die Zuverlässigkeit ab.
  - ($V2.1$) Mit zunehmender inferentieller Dichte nimmt die Zuverlässigkeit zu. Gegeben eine Evidenzmenge mit falschen Aussagen gilt $V2.1$ für Absolut-Bestätigung nur in Abhängigkeit von der Größe der Evidenzmenge.
  - ($V2.2$) Mit zunehmender Größe der Evidenzmenge nimmt die Zuverlässigkeit zu. Gegeben eine Evidenzmenge mit falschen Aussagen, gilt $V2.2$ nur in Abhängigkeit von der inferentiellen Dichte. Es gibt jedoch zwei Ausnahmen. Erstens, gegeben eine Evidenzmenge mit falschen Aussagen gilt $V2.2$ für eine bestimmte Form der Relevanz-Bestätigung, unabhängig von der inferentiellen Dichte. Zweitens, gegeben eine Evidenzmenge mit ausreichend großer Anzahl falscher Aussagen gilt $V2.2$ nicht für Absolut-Bestätigung, unabhängig von der inferentiellen Dichte.
  - ($V2.3$) Mit zunehmender Größe der Evidenzmenge nehmen die Unterschiede bezüglich der Zuverlässigkeit zwischen Absolut- und Relevanz-Bestätigung ab.
  - ($V2.4$) Mit zunehmender inferentieller Dichte nehmen die Unterschiede bezüglich der Zuverlässigkeit zwischen Absolut- und Relevanz-Bestätigung ab.

- ($V3$) Situationen in denen der Bestätigungsgrad einer Hypothese Evidenz ist für ihre Wahrheit. Es gibt Situationen, in denen Bestätigung ein zuverlässiger veritistischer Indikator für die Wahrheit einer Hypothese ist. Es zeigt sich, dass Absolut-Bestätigung häufiger ein zuverlässiger veritistischer Indikator ist als Relevanz-Bestätigung. Für einen ausreichend kleinen Anteil an wahren Aussagen in der Evidenzmenge ist Bestätigung kein zuverlässiger veritistischer Indikator, unabhängig vom verwendeten Bestätigungsmaß.
- ($V4$) Ausnahmen von der epistemischen Regel Überzeugungen anhand des Konzepts der Bestätigung zu bilden. Es gibt Situationen, in denen ist es nicht vernünftig Überzeugungen anhand des Konzepts der Bestätigung zu bilden.

Dies sind gerade jene, in denen Bestätigung kein zuverlässiger Indikator für die Wahrheit einer Hypothese ist.

Der zweite Teil meiner Arbeit, das heißt Kapitel 3, rekonstruiert und analysiert eine historische wissenschaftliche Debatte, nämlich die great Devonian controversy. Deren Teilnehmer schließen anhand von sich oftmals ändernden Evidenzmengen auf unterschiedliche Datierungen der älteren Gesteinsschichten Devons. Teil dieser Schlüsse sind sogenannte mineralogische bzw. fossile Kriterien, die den mineralogischen Charakter bzw. Fossiliengehalt einer bestimmten Gesteinsschicht mit einem bestimmten geologischen Zeitalter verknüpfen.

Grundlage meiner Rekonstruktion ist die Theorie dialektischer Strukturen. Auf diese Art und Weise gelingt es dieser Arbeit inferentielle Beziehungen zwischen Hypothesen und Evidenzen aufzudecken. Meine argumentationstheoretische Rekonstruktion der Debatte illustriert verschiedene wissenschaftsphilosophische Konzepte:

- ($H1.1$) Für jede empirische Aussage der Debatte gilt: Es gibt eine Abhängigkeit zwischen ihrem theoretischen Kontext und ihrer vernünftigen Akzeptierbarkeit. Somit illustriert die Debatte das Konzept der Theoriegeladenheit. Unter heutigen Wissenschaftsphilosophen ist dieses Konzept recht unkontrovers, vergleiche zum Beispiel (Boyd and Bogen, 2021).
- ($H1.2$) Für die Debatte gilt, dass ein mineralogisches bzw. fossiles Kriterium eine empirische Aussage nur unter Berücksichtigung weiterer Annahmen, sogenannter Hilfsannahmen, impliziert. Somit illustriert sie die Unbestimmtheitsthese von Duhem, vergleiche (Duhem, 1954).
- ($H1.3$) Es existiert eine Vielfalt an mineralogischen bzw. fossilen Kriterien. Diese sind die meiste Zeit höchst umstritten. Nur am Ende der Debatte können sich alle Teilnehmer auf ein Kriterium einigen. Die Debatte illustriert auf diese Weise das Ringen um standardisierte Methoden zur Gewinnung empirischer Aussagen.

Die Debatte endet mit einem Konsens zwischen den Hauptteilnehmern. Dieser beinhaltete nicht nur eine Einigung über die Datierung der älteren Gesteinsschichten in Devon, sondern auch eine Einigung über viele Evidenzen. Wie änderten die Teilnehmer ihre Überzeugungen so, dass sie letztlich einen Konsens erzielten?

Worin besteht der Konsens und wie ändert er sich im Laufe der Zeit? Bramson et al. (2017) bemessen Konsens bzw. Polarisierung anhand von Gruppen und verschiedensten Polarisationsmaßen. Aus praktischen Gründen fokussiert sich meine

Analyse der Debatte auf zwei Polarisationsmaße, nämlich Gruppenanzahl und Gruppengröße. Durch die Anwendung eines Ähnlichkeitsmaßes zur endogenen Definition von Gruppen werden jedoch auch zwei weitere Polarisationsmaße bestimmt, nämlich die Ähnlichkeit der Überzeugungen innerhalb derselben Gruppe bzw. verschiedener Gruppen. Für jede Phase der Debatte werden Gruppen endogen durch die Ähnlichkeit von Datierungshypothesen bzw. Evidenzmengen bestimmt. Zusätzlich werden, für jede Phase der Debatte, Gruppen exogen durch die Akzeptanz gewisser Evidenz bestimmt. Beide Ansätze verfeinern und erweitern die Polarisationsanalyse von Rudwick (1988). Im Folgenden sind die wichtigsten Resultate meiner Polarisationsanalyse aufgelistet.

- ($H2.1$) Die in meiner Arbeit exogen definierten Gruppen sind niemals dieselben wie jene in (Rudwick, 1988). Dieselben Gruppen wie in (Rudwick, 1988) ergeben sich nur durch die Zusammenfassung von Teilnehmern mit maximal ähnlichen, das heißt gleichen, Datierungshypothesen.
- ($H2.2$) Die Ähnlichkeitsspektra von Datierungshypothesen einerseits und Evidenzmengen andererseits sind zwar ziemlich ähnlich, nämlich $[0.60, 1.00]$ und $[0.54, 0.99]$. Jedoch sind die zugehörigen Ähnlichkeitsdynamiken sehr unterschiedlich und nicht korreliert.
  - Immer, außer während des Mittelteils, akzeptieren zwei Teilnehmer dieselbe Datierungshypothese. Niemals, nicht einmal am Ende der Debatte, akzeptieren zwei Teilnehmer dieselbe Evidenzmenge.
  - Mehrere Male akzeptieren zwei Teilnehmer zur selben Zeit bemerkenswert ähnliche Evidenzmengen und unähnliche Datierungshypothesen bzw. ähnliche Datierungshypothesen und unähnliche Evidenzmengen.
- ($H2.3$) Die mittlere Ähnlichkeit ist am Ende maximal, nicht zuletzt aufgrund des Austauschs von Argumenten. Insoweit Datierungshypothesen und Evidenzmengen zusammen ein Paradigma konstituieren, illustriert dieses Resultat (Kuhn, 1983). Dort wird behauptet, dass Debatten nicht nur durch interparadigmatischen Austausch von Argumenten ausgelöst, sondern auch beendet werden.

Gibt es eine Verbindung zwischen dem Konzept der Bestätigung und den individuellen Überzeugungsänderungen der Teilnehmer der Debatte? Der Begriff der Bestätigung wird auf drei verschiedene Weisen ausbuchstabiert, nämlich durch die bereits aus dem ersten Teil meiner Arbeit bekannten Bestätigungsmaße $DOJ(h|e)$,

$Z_{DOJ}(h, e)$ und $F_{DOJ}(h, e)$. Für jeden Zeitschritt und jeden Teilnehmer lässt sich also auf dreifache Weise das Maß berechnen, mit der seine Datierungshypothese, durch die von ihm akzeptierte Evidenz, bestätigt wird. Letztendlich erhält man somit für jeden Teilnehmer drei Bestätigungsdynamiken. In Bezug auf diese lässt sich Folgendes feststellen:

- ($H1.4$) Am Anfang sowie am Ende der Debatte akzeptieren alle Teilnehmer eine maximal bestätigte Datierungshypothese (relativ zur jeweils akzeptierten Evidenz), dass heißt eine Datierungshypothese mit Bestätigungsgrad 1. Dieses Resultat ist unabhängig von der Verwendung eines bestimmten Bestätigungsmaßes.
- ($H1.5$) Für die meisten Teilnehmer und Zeitpunkte gilt: $DOJ(h|e)$ und $Z_{DOJ}(h, e)$ sind sich sowohl in ihren absoluten Werten als auch in ihren relativen Änderungen sehr ähnlich und viel kleiner als 1.
- ($H1.6$) Für die meisten Teilnehmer und Zeitpunkte gilt: $DOJ(h|e)$ und $F_{DOJ}(h, e)$ sind sich sowohl in ihren absoluten Werten als auch in ihren relativen Änderungen sehr unähnlich. Für die meisten Teilnehmer und Zeitpunkte gilt: $F_{DOJ}(h, e)$ wächst mit wachsendem $\frac{DOJ(h|e)}{DOJ(h)}$ und ist annähernd 1.

Wie ändern die Teilnehmer der Debatte ihre individuellen Überzeugungen? Meine Antwort auf diese Frage greift verschiedene Konzepte der Wissenschaftsphilosophie auf:

- ($H3.1$) Teilnehmer der Debatte ändern nicht nur ihre Datierungshypothesen, sondern auch ihre Evidenzmenge. Oftmals halten sie an einer bestimmten Datierungshypothese fest und ändern lediglich ihre Evidenzmenge. Unter der Annahme das Verhalten sei gleichwohl rational, widerspricht dieses Resultat einem strengen Falsifikationismus im Sinne von Popper (1935).
- ($H3.2$) Manche Teilnehmer halten sehr ausdauernd an partiellen Datierungshypothesen oder bestimmten Evidenzen fest bzw. geben diese nur sehr widerstrebend auf. Diese partiellen Datierungshypothesen und Evidenzen können als harter Kern ihrer Überzeugungssysteme im Sinne von Lakatos (1970) betrachtet werden.
- ($H3.3$) Datierungshypothesen und Evidenzmengen werden zumeist nur geringfügig geändert. Dieses Ergebnis illustriert (Laudan, 1984). Dort wird behauptet, dass Überzeugungssysteme nicht als Ganzes überarbeitet werden, sondern vielmehr auf eine stückweise und zögerliche Art.

Gibt es einen Zusammenhang zwischen vernünftigen Überzeugungsänderungen und dem Konzept der Bestätigung? In dieser Arbeit werden die beiden folgenden Prinzipien für vernünftige Überzeugungsänderungen auf ihre Anwendbarkeit im Falle der von mir rekonstruierten Debatte hin untersucht.

- ($RAT1$) Der Wechsel von einem Paar von Datierungshypothese und Evidenzmenge zu einem anderen ist nur dann vernünftig, falls er den Bestätigungsgrad der Hypothese nicht verringert.
- ($RAT2$) Der Wechsel von einem Paar von Datierungshypothese und Evidenzmenge zu einem anderen ist nur dann vernünftig, falls er den Bestätigungsgrad der Hypothese relativ maximiert.

In dieser Arbeit gilt per Definition: Eine Person maximiert den Bestätigungsgrad der Hypothese relativ genau dann, wenn sie ihre Datierungshypothese so wählt, dass, gegeben ihre Evidenzmenge, keine andere Datierungshypothese einen höheren Bestätigungsgrad aufweist. Für die von mir rekonstruierte Debatte zeigt sich Folgendes:

- ($H4.1$) Die meiste Zeit sind die individuellen Überzeugungsänderungen vernünftig im Sinne beider Prinzipien. Jedoch gibt es Zeitpunkte zu denen manche Überzeugungsänderungen weder vernünftig im Sinne des einen, noch im Sinne des anderen Prinzips sind. Individuelle Überzeugungsänderungen sind im Sinne beider Prinzipien am häufigsten vernünftig, falls $F_{DOJ}(h, e)$ verwendet wird.
- ($H4.2$) Geteilte Überzeugungsänderungen sind seltener vernünftig als individuelle Überzeugungsänderungen. Dies gilt sowohl für Vernunft im Sinne des einen wie auch des anderen Prinzips. Geteilte Überzeugungsänderungen sind im Sinne beider Prinzipien am häufigsten vernünftig, falls $F_{DOJ}(h, e)$ verwendet wird.

Es sei angenommen Teilnehmer der Debatte seien stets vernünftig, dann folgt aus $H4.1$, dass die oben genannten Prinzipien für vernünftige Überzeugungsänderungen nicht immer gelten. Dieses Resultat ist im Sinne von Feyerabend (1976). welcher bestreitet, dass es ausnahmslos gültige wissenschaftliche Regeln gibt. Jedoch zeigt meine Arbeit nicht, dass es überhaupt keine Regeln gibt.

Unterscheiden sich die Wege zum finalen Konsens der einzelnen Teilnehmer? Gibt es bemerkenswerte Ähnlichkeiten? Ein herausstehendes Merkmal meiner Analyse der

Konsensfindung ist die Trennung hinsichtlich Datierungshypothesen und Evidenzmengen. Dies ist eine Verfeinerung der korrespondierenden Analyse in (Rudwick, 1988). Es zeigt sich, dass für alle Teilnehmer gilt:

- ($H5.1$) Phasen der Annäherung wechseln mit Phasen der Abkehr vom finalen Konsens. Dieses Resultat ist im Sinne von Betz (2013). Dort finden sich Konsensdynamiken von kontroversen Debatten auf der Grundlage von Multi-Agenten-Simulationen.
- ($H5.2$) Eine Annäherung bzw. Abkehr von der finalen Datierungshypothese bedingt nicht notwendig eine Annäherung bzw. Abkehr von der letztlich geteilten Evidenzmenge, und umgekehrt. Dieses Resultat erweitert und verfeinert die Analyse der Konsensbildung in (Rudwick, 1988).

Verringert eine individuelle Überzeugungsänderung die Ähnlichkeit mit dem finalen Konsens genau dann, wenn diese nicht vernünftig ist?

Hier wird wiederum von denselben beiden Prinzipien für vernünftige Überzeugungsänderungen Gebrauch gemacht. Meine Analyse zeigt, dass nicht gilt: Eine individuelle Überzeugungsänderung verringert die Ähnlichkeit mit dem finalen Konsens genau dann, wenn sie den Bestätigungsgrad einer Datierungshypothese verringert. Das gilt für alle drei hier besprochenen Bestätigungsmaße. Jedoch zeigt meine Analyse, dass nach ausreichend vielen Überzeugungsänderungen, die den Bestätigungsgrad verringern, häufig eine Überzeugungsänderung folgt, die nicht nur den Bestätigungsgrad vergrößert, sondern zusätzlich auch die Ähnlichkeit mit dem finalen Konsens. Eine Serie von Abnahmen des Bestätigungsgrades scheint also ein guter Grund zu sein um die eigenen Überzeugungen beträchtlich zu ändern. Es gilt jedoch zu beachten, dass es häufig auch gute Gründe gibt an den eigenen Überzeugungen festzuhalten, trotz abnehmendem Bestätigungsgrad der Datierungshypothese. Meine Analyse zeigt des Weiteren, dass nicht gilt: Eine Überzeugungsänderung verringert die Nähe zum finalen Konsens genau dann, wenn sie den Bestätigungsgrad der Datierungshypothese nicht relativ maximiert. Dieses Ergebnis gilt unabhängig von der Verwendung eines bestimmten Bestätigungsmaßes. Weiters gilt nicht, dass eine Überzeugungsänderung die Nähe zum finalen Konsens vergrößert, wenn sie den Bestätigungsgrad relativ maximiert.

Ist die relative Maximierung des Bestätigungsgrades der Datierungshypothese eine geeignete Methode um sich dem finalen Konsens zu nähern? Gilt dies ausnahmslos? Falls nicht, können die Ausnahmen unter Verwendung des Konzepts der Evidenz

höherer Ordnung charakterisiert werden? Im Rahmen meiner Arbeit sind Beispiele für eine Evidenz höherer Ordnung die inferentielle Dichte der dialektischen Struktur zu einem gewissen Zeitpunkt und, für jeden Debattenteilnehmer, die Größe der Evidenzmenge sowie ihre Korrektheit. Im Falle der von mir rekonstruierten Debatte wird die Korrektheit der Evidenzmenge bemessen anhand der Ähnlichkeit mit dem finalen Konsens.

Für jeden Teilnehmer und jeden Zeitpunkt wird die relative Häufigkeit bestimmt, mit der eine Datierungshypothese, die den Bestätigungsgrad relativ maximiert, zusätzlich dem finalen Konsens ausreichend ähnlich ist. Es gibt einige bemerkenswerte Erkenntnisse bezüglich dieser relativen Häufigkeit:

- ($H5.3$) Die relative Häufigkeit unterscheidet sich erheblich für unterschiedliche Personen, das heißt sie ist abhängig von der Evidenzmenge. Diese Abhängigkeit kann jedoch nicht erklärt werden durch die Nähe zur finalen Evidenzmenge. Eine bestimmte Nähe ist weder eine hinreichende noch notwendige Bedingung für eine relative Häufigkeit größer 0.5. Daraus schließe ich, dass manche Evidenzen mehr Einfluss haben als andere.
- ($H5.4$) Die relative Häufigkeit unterscheidet sich erheblich für unterschiedliche Bestätigungsmaße. Um die relative Häufigkeit zu steigern, ist $Z_{DOJ}(h, e)$ den beiden anderen Bestätigungsmaßen vorzuziehen.
- ($H5.5$) Vor dem Zeitpunkt $4a$ gilt unabhängig von einem bestimmten Bestätigungsmaß und einer bestimmten Person: Die relative Häufigkeit ist kleiner oder gleich 0.5. Daraus folgere ich, dass, zu Beginn der Debatte, die Methode der relativen Maximierung des Bestätigungsgrades nicht gut geeignet ist um sich dem finalen Konsens anzunähern, und zwar unabhängig von einer bestimmten Person.
- ($H5.6$) Vom Zeitpunkt $7b$ an bis zum Ende gilt für alle Bestätigungsmaße und die meisten Teilnehmer: Die relative Häufigkeit ist gleich 1. Am Ende ist dies für alle Teilnehmer der Fall. Nach einer Anhäufung von Argumenten und Evidenzen sowie wiederholten Überzeugungsänderungen der Teilnehmer, schließt die Debatte also zu einem Zeitpunkt an dem die relative Maximierung des Bestätigungsgrades eine geeignete Methode ist sich dem finalen Konsens zu nähern, und zwar unabhängig von einer bestimmten Person.

Die zwei letztgenannten Punkte illustrieren ein Modell der wissenschaftlichen Konsensfindung, das sowohl Ideen des Rationalismus als auch Anti-Rationalismus bein-

haltet, nämlich das sogenannte Kompromissmodell von Kitcher (1993). Es sei angenommen, dass (*i*) die Teilnehmer der Debatte den Bestätigungsgrad ihrer Datierungshypothesen relativ maximieren und (*ii*) kognitiver Fortschritt anhand der Annäherung an den finalen Konsens bemessen wird. In diesem Fall illustriert $H5.5$ Bedingung $C4$ des Kompromissmodells, das heißt die Aussage, dass, zu Beginn einer wissenschaftlichen Debatte, die kognitiven Prozesse der letztlichen Sieger nicht fortschrittlicher sind als diejenigen der letztlichen Verlierer. Unter denselben zwei Annahmen illustriert $H5.6$ Bedingung $C5$ des Kompromissmodells. Grob gesagt behauptet diese, dass wissenschaftliche Debatten ein Ende finden, wenn, als Ergebnis von Argumentation, Evidenzanhäufung und Überzeugungsänderungen, ein bestimmter kognitiver Prozess, der allen Teilnehmern möglich ist, fortschrittlicher ist als alle anderen.

Summary

The overall motivation of this study arises from two assumptions about science which are quite uncontroversial in today's philosophical discussions: Firstly, it is assumed that empirical evidence is fallible and many of our present or past evidential beliefs have actually been false. Secondly, it is assumed that our ability to correctly infer the truth or falsity of a hypothesis depends on whether our body of evidence is correct. Departing from these two basic assumptions, the following question arises: How can scientists reason with false evidence and reliably infer hypotheses?

In real scientific debates, beliefs are often formed according to some notion of relevance confirmation. Relevance confirmation belongs to the realm of Bayesian confirmation theory, which is a probabilistic theory of confirmation. For relevance confirmation, the degree of confirmation, that a hypothesis receives from some evidence, depends on the absolute probability of the hypothesis and the conditional probability of the hypothesis given the evidence. Another form of confirmation is absolute confirmation, where the degree of confirmation only depends on the conditional probability of the hypothesis given the evidence. There are multitudes of discussed Bayesian confirmation measures. Only for three of them, it holds: There are good reasons to consider confirmation as an extension of the concept of deductive entailment. Here, I take it that good reasons are those given in (Crupi and Tentori, 2013) and (Crupi et al., 2007). These three confirmation measures are the conditional probability of the hypothesis given the evidence, $P(h|e)$, $Z_P(h, e)$ and $F_P(h, e)$. The two latter ones are relevance confirmation measures, firstly defined by Crupi and Tentori (2010) and Branden Fitelson (2004) heavily relying on (Kemeny and Oppenheim, 1952). The first one is an absolute confirmation measure.

In this thesis, a dialectic account of probability is used. As Betz (2010) shows, the degree of justification of a hypothesis given some evidence and a dialectical structure, $D(h|e)$, satisfies the Kolmogorov axioms. Therefore, it is a probability. A dialectical structure consists of sentences and deductive valid arguments possibly attacking or

supporting one another. According to the theory of dialectical structures, every state of a debate can be represented by a dialectical structure and the positions of its proponents. The theory of dialectical structures is a formal model of complex argumentation and developed in (Betz, 2010).

Forming beliefs according to confirmation, is this a valuable way of reasoning with false evidence? What different ways of forming beliefs according to confirmation are there and do they differ in their truth-conduciveness? How truth-conducive is it to adjust one's beliefs to degrees of confirmation? How truth-conducive is it to only accept those hypotheses which are sufficiently confirmed by one's evidence? These questions trigger a *statistical turn* and are answered in detail in part one of this thesis, that is chapter 2.

What is going on in real scientific debates? How do participants change their beliefs such that they finally reach a consensus? Do they shift from one group of a dating hypothesis and evidential beliefs to another such that their degrees of confirmation increase? Do they shift from one group of a dating hypothesis and evidential beliefs to another such that there is no other hypothesis which is better confirmed? Do such shiftings always increase similarity with the final consensus? These questions trigger a *historic turn* and are answered in detail in part two of this thesis, that is chapter 3.

In both parts of this thesis, it holds that analyzing debates in terms of confirmation needs the help of computational techniques and resources. In doing so, this thesis contributes to the program of computational philosophy. According to Grim and Singer (2020), computational philosophy comprises all philosophical research making use of computational techniques. In both parts of my thesis, debates are represented as dialectical structures and analyzed in terms of confirmation relying on justification, which is another concept of the theory of dialectical structures. An important benefit of this theory is its connectivity to computational analyses since a dialectical structure is a Boolean formula.

Part one of this thesis makes use of 1000 simulated dialectical structures drawn from (Betz, 2013) as well as Monte-Carlo techniques. In part two of this thesis, Argdown as developed by Christian Voigt (2018), is used to implement and output dialectal structures corresponding to states of a historic debate. In both parts of my thesis, calculating degrees of justification needs computational techniques and resources. For all dialectical structures of this thesis, it holds: Calculating a

hypothesis' degree of justification is beyond human capability. Computational analyses are performed using the computer algebra system Mathematica from Wolfram Research, Inc. (2019) and computing resources from Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT) (2017) and KIT's Department of Philosophy, ITZ and ITAS (2017). All of my coding, that is analyses programs and dialectical structures of my reconstruction of a historic debate, is accessible via https://github.com/cguenth/CONFasVI.git and https://github.com/cguenth/GDC.git.

In both parts of my thesis, it shows that there are (*i*) differences between absolute and relevance confirmation and (*ii*) no rules without exceptions. The latter result confirms Feyerabend (1976) stating that there is no scientific rule without any exceptions. However, it does not support relativism in the sense of "anything goes". In both parts of my thesis, exceptions are characterized in terms of higher-order evidence.

What is higher-order evidence? First, I present my own notion of higher-order evidence. I consider a body of evidence as first-order evidence and a statement about first-order evidence as higher-order evidence. Examples are statements about the amount and correctness of first-order evidence, the argumentative role of first-order evidence as well as the properties of the argumentative structure into which first-order evidence is embedded. In this thesis, analyses are performed for three examples of higher-order evidence, namely the inferential density of the dialectical structure at a certain time step, and, for every person, the amount of evidence claims and the ratio of true evidence claims.

Among today's epistemologists, there is a lively debate about the nature of higher-order evidence. It is characterized at least in two different ways: Higher-order evidence is considered as evidence about (*i*) the character of the evidence or (*ii*) an agent's capacities for responding rationally to the evidence. According to Talbott (2016b), higher-order evidence is defeating evidence of a certain kind. It is a certain kind of undercutting defeater, namely a reliability defeater as described in (Pollock, 1984).

Second, I connect higher-order evidence in the sense of Talbott (2016b) with my own notion of higher-order evidence. Presuppose that participants of a debate form their beliefs according to confirmation. In this case, higher-order evidence in the sense of Talbott (2016b) is everything defeating the reliability of this cognitive process. Betz (2015) has shown that higher-order evidence in my sense, namely inferential

density and the amount of evidence, allows us to estimate the reliability of absolute confirmation as a veritistic indicator for the truth of a hypothesis. However, these results are limited. First, Betz (2015) assumes a totally true body of evidence. Second, Betz (2015) only considers absolute confirmation as a veritistic indicator. The first part of my thesis expands upon (Betz, 2015) not only by considering (i) partly incorrect bodies of evidence but also (ii) relevance confirmation measures.

In this thesis, the reliability of a veritistic indicator is assessed via a statistical hypothesis test based on Monte-Carlo simulations. The reliability of a veritistic indicator improves, if significance decreases and power increases. The hypothesis $\neg h$ is tested against the alternative hypothesis h. For every statistical test, there is a critical region, that is a region where $\neg h$ has to be rejected, which is chosen such that the significance equals 0.05. Hence, this thesis expands upon (Betz, 2015) by assessing the reliability of confirmation in a new way.

Today, there are several re-evaluating principles stating that I have to re-evaluate my former beliefs in light of certain evidence about the process producing these beliefs. Part one of my thesis underpins certain re-evaluating principles, as for example the $MERF$ principle as introduced by Talbott (2016b) and the integration principle as introduced by Christensen (2008), in a twofold way. First, it assesses the reliability of special cognitive processes, namely certain modes of belief formation according to confirmation. Second, it identifies reliability-relevant categorizations of these processes in terms of higher-order evidence. According to the integration principle, object-level beliefs must reflect meta-level beliefs about the reliability of the cognitive process. According to the $MERF$ principle, a belief must be revised, if it *does not* equal the reliability of the cognitive process (unless there is some categorization of this process such that it *does* equal the reliability of the cognitive process, at least approximately). According to Talbott (2016b), the reliability of a cognitive process is given by the expected relative frequency of truths among beliefs which are produced by this very process.

Among today's epistemologists, there is also a lively debate about the bearing of higher-order evidence. For example, is there a situation in which evidence of evidence (for some hypothesis h) is itself evidence for h? This thesis is claimed by some epistemologists, see for example (Feldman, 2005). However, it is also contested by some others, see for example (Fitelson, 2012). Presuppose that confirmation is another example of higher-order evidence and there is a situation in which confirmation is a

reliable veritistic indicator. Then, there is a situation in which evidence of evidence (for some hypothesis h) is itself evidence for h.

The main findings of part one of my thesis are:

- ($V1$) The relative frequency of truths among hypotheses with a certain degree of confirmation. It shows that, for a totally correct body of evidence, higher-order evidence influences the relative frequency, only if using a relevance confirmation measure.
- ($V2$) The reliability of confirmation as a veritistic indicator. Independent of higher-order evidence, it shows that absolute confirmation is a more reliable indicator than relevance confirmation. Higher-order evidence influences the reliability of confirmation as a veritistic indicator and there are differences between absolute and relevance confirmation.
  - ($V2.0$) As the ratio of false evidence claims increases, the reliability decreases.
  - ($V2.1$) As the inferential density increases, the reliability increases. For a body of evidence including false evidence claims and absolute confirmation, the truth of $V2.1$ depends on the amount of evidence.
  - ($V2.2$) As more and more evidence is accumulated, the reliability increases. For a body of evidence including false evidence claims, the truth of $V2.2$ depends on the inferential density. However, there are exceptions. First, for a body of evidence including false evidence claims and a certain kind of relevance confirmation, $V2.2$ holds, independent of the inferential density. Second, for a sufficiently large amount of false evidence claims and absolute confirmation, $V2.2$ does not hold, independent of the inferential density.
  - ($V2.3$) As more and more evidence is accumulated, differences in reliability between absolute and relevance confirmation decrease.
  - ($V2.4$) As the inferential density increases, differences in reliability between absolute and relevance confirmation decrease.
- ($V3$) Situations in which its degree of confirmation is evidence for the confirmed hypothesis. There are situations where its degree of confirmation is a reliable indicator for the truth of a hypothesis. It shows that absolute confirmation is more often a reliable veritistic indicator than relevance confirmation.

For a sufficiently small amount of true evidence claims, confirmation is no reliable veritistic indicator, independent of a certain confirmation measure.

- ($V4$) Exceptions to the epistemic rule of forming beliefs according to confirmation. There are situations where it is not rational to form beliefs according to confirmation, namely those where confirmation is no reliable indicator for the truth of a hypothesis.

Part two of my thesis, that is chapter 3, reconstructs a historic scientific debate, namely the great Devonian controversy. During this debate, participants infer different hypotheses about the age of all the older strata in Devonshire from evidential beliefs often changing. These inferences are based on so-called mineralogical and fossil criteria, connecting the mineralogical character and fossil content of certain strata with certain geological ages, respectively.

My reconstruction of the debate relies on the theory of dialectical structures. In doing so, it reveals relations between evidence and hypotheses. There are several concepts of philosophy of science which are illustrated by my reconstruction of the great Devonian controversy:

- ($H1.1$) For all empirical statements of the great Devonian controversy, it holds: There is a dependence between its theoretical context and rational acceptance. Hence, the debate illustrates nicely the concept of theory-ladenness, which is uncontroversial in today's philosophy of science, see for example (Boyd and Bogen, 2021).
- ($H1.2$) For the great Devonian controversy, it shows that an empirical statement is implied by some mineralogical or fossil criterion, only if it is conjoined with some auxiliary assumptions. Therefore, the debate illustrates Duhemian underdetermination, compare (Duhem, 1954).
- ($H1.3$) There are several criteria and most of them are highly controversial most of the time. Only at the end, there is a criterion which all participants agree upon. Therefore, the great Devonian controversy illustrates the struggle about standardizing methodological rules for generating empirical statements.

For the great Devonian controversy, there is finally a consensus between the main participants, not only regarding the dating of all the older strata but also most of the evidential statements. How did participants change their beliefs such that they finally reach a consensus?

What is consensus? How does consensus change with time? Bramson et al. (2017) assess consensus respectively polarization in terms of groups and propose a variety of polarization measures. For practical reasons, my analyses of the great Devonian controversy center on community fragmentation and size parity. However, introducing and using a similarity measure to define groups endogenously, group consensus as well as distinctness are assessed as well. For every time step, not only groups are identified exogenously by accepting a certain piece of evidence but endogenously using degrees of similarity between dating hypotheses respectively bodies of evidence. This way, my thesis enhances (Rudwick, 1988). The main results of these analyses are:

- ($H2.1$) For exogenous clustering, groups are never the same as those in (Rudwick, 1988). This is not true for endogenous clustering. Clustering maximally similar dating hypotheses, groups are the same as those in (Rudwick, 1988).
- ($H2.2$) Similarity spectra of dating hypotheses and bodies of evidence are quite similar, namely $[0.60, 1.00]$ and $[0.54, 0.99]$. However, similarity dynamics of dating hypotheses and bodies of evidence do not coincide:
  - Except during the middle section, there are always some persons accepting the same dating hypothesis. Never, not even at the end, there are two persons accepting the same body of evidence.
  - Several times, there are two persons accepting, at the same time, remarkably similar bodies of evidence but unsimilar dating hypotheses, and vice versa.
- ($H2.3$) The average degree of similarity is maximal at the final step, not at last due to argumentation. In so far as dating hypotheses and bodies of evidence together constitute a paradigm, this result may be understood as an illustration of Kuhn (1983) stating that controversies are not only triggered but also resolved by inter-paradigmatic exchange of arguments.

Is there some connection between individual belief change and confirmation? For the great Devonian controversy, this thesis quantifies the notion of confirmation in three different ways, namely using $DOJ(h|e)$, $Z_{DOJ}(h, e)$ and $F_{DOJ}(h, e)$. In doing so, it reveals degrees of confirmation and *confirmation dynamics*.

- ($H1.4$) The great Devonian controversy starts and ends with all main participants accepting a dating hypothesis with a maximal degree of confirmation

(given a certain evidence), that is with a degree of 1, independent of a certain confirmation measure.

- ($H1.5$) For most time steps and participants, $DOJ(h|e)$ and $Z_{DOJ}(h,e)$ are rather similar, both in value and relative changes, and much smaller than 1.
- ($H1.6$) For most time steps and participants, $DOJ(h|e)$ and $F_{DOJ}(h,e)$ are rather unsimilar, both in value and relative changes. For most time steps and participants, $F_{DOJ}(h,e)$ increases with increasing $\frac{DOJ(h|e)}{DOJ(h)}$ and is fairly approximated by 1.

How do participants of the great Devonian controversy change their beliefs individually? Answers to this question pick up several concepts of philosophy of science:

- ($H3.1$) Participants do not only change their dating hypotheses, but also their evidential beliefs. Often, participants hold on to a certain dating hypothesis while changing evidential beliefs. Given that participants are rational, this result dis-confirms strict falsificationism in the sense of Popper (1935).
- ($H3.2$) For the great Devonian controversy, there are dating hypotheses as well as evidential beliefs, which are constantly kept, or at least only very reluctantly given up. Hence, these dating hypotheses and evidential beliefs can be considered as hard-core assumptions in the sense of Lakatos (1970).
- ($H3.3$) Most of the time, dating hypotheses as well as bodies of evidence are only slightly altered. This illustrates Laudan (1984) stating that beliefs are not revised as a whole, but rather in a piecemeal and reluctant way.

Is rationality in belief change related with some kind of evidential support? In this thesis, evidential support is spelled out in terms of confirmation, that is $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. For the great Devonian controversy, do the following principles of rational belief change apply?

- ($RAT1$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it does not decrease the dating hypothesis's degree of evidential support.
- ($RAT2$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it maximizes the dating hypothesis's degree of evidential support.

Here, a person maximizes her dating hypothesis's evidential support, if she chooses a

dating hypothesis such that, given her body of evidence, there is no better confirmed dating hypothesis. For the great Devonian controversy, the following shows:

- ($H4.1$) Most of the time, individual belief changes are rational. However, there are individual belief changes which are not rational. Using $F_{DOJ}(h, e)$, individual belief changes are *more* often rational than using one of the other two confirmation measures.
- ($H4.2$) Shared belief changes are *less* often rational than individual belief changes. Using $F_{DOJ}(h, e)$, shared belief changes are *more* often rational than using one of the other two confirmation measures.

Presuppose that participants of the great Devonian controversy are rational. Together with $H4.1$ it follows that there are exceptions to the two previously introduced principles of rational belief change. They are no strict rules. This confirms Feyerabend (1976) stating that there is no scientific rule without any exceptions. However, it does not support relativism in the sense of "anything goes".

For participants of the great Devonian controversy, do roads to the final consensus differ? Are there strong similarities? A prominent point of my analysis of final consensus formation is the separation between dating hypotheses and bodies of evidence. This is a refinement of the analysis of final consensus formation given in (Rudwick, 1988). For all main participants, it holds:

- ($H5.1$) Approaching alternates with distancing the final consensus. This is in line with Betz (2013) investigating consensus-conduciveness of controversial debates by means of multi-agent simulations.
- ($H5.2$) Approaching the final consensus in terms of dating hypotheses does not imply an approachment in terms of bodies of evidence, and vice versa. This is a refinement of the analysis of consensus formation given in (Rudwick, 1988).

For the great Devonian controversy, does an individual belief change decrease similarity with the final consensus, iff it is not rational? Here as before, the same two principles of rational belief change are considered, relating rationality with evidential support. As a result of my analyses, it does not hold that a belief change distances the final consensus, iff it decreases evidential support. This is true for all three confirmation measures. However, it shows that, after a sufficiently large number of successive belief changes decreasing evidential support, there is often a considerable

change in similarity with the final consensus. Hence, successive decrease of evidential support seems to be a reason for changing beliefs. Note that there are reasons for not changing beliefs, even if evidential support decreases. As a further result of my analyses, it does not hold that a belief change distances the final consensus, iff it does not maximize evidential support. This is true for all three confirmation measures. Further, it does not hold that a belief change approaches the final consensus, if it maximizes evidential support.

Is maximizing evidential support well designed to approach the final consensus? Is this a rule with exceptions? If so, how can these exceptions be characterized? In terms of higher-order evidence? Remember that, in the first part of my thesis, analyses are performed for three examples of higher-order evidence, namely the inferential density of the dialectical structure at a certain time step, and, for every person, the amount of evidence claims and the ratio of true evidence claims. For the great Devonian controversy, the correctness of first-order evidence is assessed in terms of similarity with the final body of evidence.

For every time step and person, the ratio of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize evidential support is determined. There are several things about this ratio which should be noted:

- ($H5.3$) There are big differences between persons. Hence, there is a dependence between this ratio and a person's body of evidence. However, it cannot be assessed in terms of a body's similarity with the final consensus. A certain degree of similarity is neither a sufficient nor a necessary condition for the ratio being greater 0.5. Hence, some evidential claims seem to have more impact than others.
- ($H5.4$) There are differences between confirmation measures. In order to maximize the ratio, $Z_{DOJ}(h, e)$ is better than the other two confirmation measures.
- ($H5.5$) For all three confirmation measures and persons, before time step $4a$, it holds: The ratio is less or equal 0.5. Hence, during early phases of the debate, maximizing evidential support is not well designed to approach the final consensus, independent of a certain person.
- ($H5.6$) For all three confirmation measures and most main participants, from time step $7b$ till the end, the ratio equals 1. At the final step, this is true for all persons. Hence, the debate is closed when, as a result of argumentation,

evidence accumulation and belief changes, maximizing evidential support is maximally well designed to approach the final dating hypothesis, independent of a certain person.

The two latter results newly illustrate a model of the closure of major scientific debates embodying ideas of rationalism as well as anti-rationalism, namely the compromise model as introduced in (Kitcher, 1993). Presume that (*i*) participants of the great Devonian controversy undergo the process of maximizing evidential support and (*ii*) cognitive progress is considered as approaching the final dating hypothesis. Then, $H5.5$ confirms condition $C4$ of the compromise model stating that, "[d]uring early phases of scientific debate, the processes undergone by the ultimate victors are (usually) no more well designed for promoting cognitive progress than those undergone by the ultimate losers" (Kitcher, 1993, p. 201). Further, presuming the same two assumptions, $H5.6$ confirms condition $C5$ of the compromise model, roughly stating that scientific debates end when, as a result of argumentation, evidence accumulation and belief changes, a certain cognitive process, which is executable for all participants, performs better in terms of cognitive progress than all the others undergone by participants of the debate.

Nomenclature

Confirmation

$CONF_{DOJ}(h, e)$ Some probabilistic confirmation measure assessing absolute or relevance confirmtaion, relying on a dialectical concept of probability, namely justification, page 58

$DOJ(h \mid e)$ or DOJ A hypothesis's degree of justification given some body of evidence. For the great Devonian controversy, h is always a dating of all the older strata in Devon. Further, it is a probabilistic confirmation measure assessing absolute confirmation, page 51

$F_{DOJ}(h, e)$ or F A certain probabilistic confirmation measure assessing relevance confirmation, relying on a dialectical concept of probability, namely justification, page 58

$Z_{DOJ}(h, e)$ or Z A certain probabilistic confirmation measure assessing relevance confirmation, relying on a dialectical concept of probability, namely justification, page 58

Veritistically Valuable Reasoning

$\mid e_T \mid / \mid e \mid$ Ratio of true evidence claims, one example of higher-order evidence, page 60

α Significance of a statistic hypothesis test, page 50

$\mid e \mid$ Size of the body of evidence, one example of higher-order evidence, page 60

ω Critical region of a statistic hypothesis test, page 50

$D(\tau)$ Inferential density of a dialectical structure, one example of higher-order evidence, page 51

HOE Higher-order evidence, that is a statement about first-order evidence, page 60

$RAT0$ A certain epistemic rule about forming beliefs according to confirmation, page 69

1-β Power of a statistic hypothesis test, page 50

Historic Scientific Reasoning

$B_1 ... B_5$ 5 sentences, so-called interpretative boundaries, used for exogenously defining groups of participants of the great Devonian controversy, page 105

CM, ML, ORS, SIL and CAM 5 geological ages in descending order, namely Coal Measures, Mountain Limestone, Old Red Sandstone, Silurian and Cambrian

CON Similarity with the final consensus. Using the similarity measure SIM, similartiy between a person's dating hypothesis and the dating hypothesis of the final consensus (SIM_h) respectively the similarity between a person's body of evidence and the body of evidence of the final consensus (SIM_e) is assessed, page 142

GDC An exceptionally well documented scientific debate among 19th century geologists about the dating of all the older strata in Devonshire, page 77

MC, BCL and NC The main part of the Culm strata, the black Culm limestone and the Non-Culm strata, constituting a partition of all the older strata in Devon

$RAT1$ A certain epistemic rule about forming beliefs according to confirmation, page 128

$RAT2$ A certain epistemic rule about forming beliefs according to confirmation, page 128

SIM Similarity measure, assessing the similarity between two dating hypotheses respectively two bodies of evidence. It is used for endogenously defining groups of participants of the great Devonain controversy as well as assessing the similarity between a participant's beliefs and the final consensus, page 110

t_a and t_b with $t \in \{0, ..., 8\}$ Time steps of the great Devonian controversy. Each of the 9 time steps is splitted a second time, separating shared from individual belief changes, page 99

DLB,MUR,LYE,PHI,SED,AUS Six main participants of the great Devonian controversy, namely Henry de la Beche, Roderick Murchison, Charles Lyell, John Phillips, Adam Sedgwick and Robert Austen

1. Introduction

1.1. Motivation and Points of Reference

Science is an epistemic authority. According to Goldman (2003), there are at least two reasons. First, science is very successful in explaining and predicting phenomena. Second, there is no other human practice, which performs better.

The overall motivation of this study arises from two assumptions about science, which are quite uncontroversial in today's philosophical discussions: Firstly, it is assumed that empirical evidence is fallible and many of our present or past evidential beliefs have actually been false. Secondly, it is assumed that our ability to correctly infer the truth or falsity of a hypothesis depends on whether our body of evidence is correct. Departing from these two basic assumptions, the following question arises: How can scientists reason with false evidence and reliably infer hypotheses?

In real scientific debates, beliefs are often formed according to some notion of confirmation. Forming beliefs according to confirmation, is this a valuable way of reasoning with false evidence? What does "valuable" mean? Are scientists striving for truth? Are our best scientific theories true? In history of science, there are various examples of predictively and explanatorily successful theories, which were false, that is, at least, some of their metaphysical assumptions were false. Think for example of the caloric theory or the phlogiston theory of combustion. Hence, the truth of an inferred hypothesis stays questionable, even if it is the result of scientific reasoning. So instead of truth, sometimes one has to focus on consensus.

What different ways of forming beliefs according to confirmation are there and do they differ in their truth-conduciveness? How truth-conducive is it to adjust one's beliefs to degrees of confirmation? How truth-conducive is it to only accept those hypotheses which are sufficiently confirmed by one's evidence? These questions trigger a *statistical turn* and are answered in detail in part one of this thesis, that is chapter 2.

What is going on in real scientific debates? How do participants change their beliefs such that they finally reach a consensus? Do they shift from one group of a dating hypothesis and evidential beliefs to another such that their degrees of confirmation increase? Do they shift from one pair of hypothesis and body of evidence to another such that there is no other hypothesis which is better confirmed? Do such shiftings always increase similarity with the final consensus? These questions trigger a *historic turn* and are answered in detail in part two of this thesis, that is chapter 3.

In both parts of my thesis, it holds that analyzing debates in terms of confirmation needs the help of computational techniques and resources. In doing so, this analysis contributes to the program of computational philosophy. According to Grim and Singer (2020), computational philosophy comprises all philosophical research making use of computational techniques. In this thesis, computational analyses are performed using the computer algebra system Mathematica from Wolfram Research, Inc. (2019) and computing resources from Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT) (2017) and KIT's Department of Philosophy, ITZ and ITAS (2017). All of my coding, that is analysis programs and dialectical structures of a historic debate, are accessible via https://github.com/cguenth/CONFasVI.git and https://github.com/cguenth/GDC.git.

In both parts of my thesis, debates are represented as dialectical structures and analysed in terms of confirmation relying on justification, which is another concept of the theory of dialectical structures. For more information on confirmation and the theory of dialectical structures, see sec. 1.1.1. An important benefit of the theory of dialectical structures is its connectivity to computational analyses. This is due to the fact that a dialectical structure is a boolean formula. Part one of this thesis makes use of 1000 simulated dialectical structures drawn from Betz (2013) as well as Monte-Carlo techniques. In part two of this thesis, Argdown, as developed by Christian Voigt (2018), is used to implement and output dialectal structures corresponding to states of a historic debate. In both parts of my thesis, calculating degrees of justification needs computational techniques and resources. For the dialectical structures of this thesis, calculating a hypothesis' degree of justification is beyond human capability.

In both parts of my thesis, are there no rules without exceptions? This would confirm Feyerabend (1976) stating that there are no strict rules for scientists. However, it must not support relativism in the sense of "anything goes". How can exceptions

be characterized? In terms of higher-order evidence? For more information on the concept of higher-order evidence, see sec. 1.1.2.

Part one of my thesis, that is chapter 2, shows how higher-order evidence allows us to infer the reliability of confirmation. In doing so, it contributes to the program of veritistic social epistemology as introduced in (Goldman, 2003). For more information on reliability and veritistic social epistemology, see sec. 1.1.3.

Part two of my thesis, that is chapter 3, reconstructs a historic scientific debate, namely the great Devonian controversy. Using the theory of dialectical structures, my reconstruction reveals relations between evidence and hypotheses. Further, the reconstruction is analyzed in terms of polarization as well as confirmation, investigating the consensus-conduciveness of certain modes of belief formation according to confirmation. For more information on my reconstruction and analyses of the great Devonian controversy, see sec. 1.1.4.

1.1.1. Confirmation and Dialectical Structures

In real scientific debates, beliefs are often formed according to some notion of relevance confirmation. Relevance confirmation belongs to the realm of Bayesian confirmation theory which is a probabilistic theory of confirmation. For relevance confirmation, the degree of confirmation that a hypothesis receives from some evidence only depends on the absolute probability of the hypothesis and the conditional probability of the hypothesis given the evidence. Another form of confirmation is absolute confirmation, where the degree of confirmation only depends on the conditional probability of the hypothesis given the evidence. There are multitudes of discussed Bayesian confirmation measures. Only for three of them, it holds: There are good reasons to consider confirmation as an extension of the concept of deductive entailment. Here, I take it that good reasons are those given in (Crupi and Tentori, 2013) and (Crupi et al., 2007). These three confirmation measures are the conditional probability of the hypothesis given the evidence, $P(h|e)$, $Z_P(h, e)$ and $F_P(h, e)$. The two latter ones are relevance confirmation measures, firstly defined by Crupi and Tentori (2010) and Branden Fitelson (2004) heavily relying on (Kemeny and Oppenheim, 1952). The first one is an absolute confirmation measure.

In this thesis, a dialectic account of probability is used. As Betz (2010) shows, the degree of justification of a hypothesis given some evidence and with respect

to a dialectical structure, $D(h|e)$, satisfies the Kolmogorov axioms. Therefore, it is a probability. A dialectical structure consists of sentences and deductive valid arguments possibly attacking and supporting one another. According to the theory of dialectical structure, every state of a debate can be represented by a dialectical structure and the positions of its proponents. A complete (partial) position is represented by a truth-value assignment to all (some) sentences in the dialectical structure. A complete position is dialectically consistent, iff (*i*) it assigns complementary truth values to a sentence and its negation and (*ii*) considers conclusions of arguments with true premises as true. A hypothesis's conditional degree of justification is given by the ratio of the number of complete and dialectical consistent positions extending the hypothesis as well as the body of evidence, and the number of complete and dialectical consistent positions only extending the hypothesis. The theory of dialectical structures is a formal model of complex argumentation and developed in (Betz, 2010). Reconstructing a state of a debate as a dialectical structure reveals not only auxiliary assumptions and inferential relations, but also argument types and clusters.

There are also other theories of confirmation such as for example Hempelian confirmation and hypothetico-deductivism. However, being qualitative theories of confirmation, both theories are not suited for my analyses. For more information on these two theories of confirmation see for example (Hempel, 1945a), (Hempel, 1945b), or (Huber, 2008).

1.1.2. Higher Order Evidence

What is higher-order evidence? First, I present my own notion of higher-order evidence. I consider a body of evidence as first-order evidence and a statement about first-order evidence as higher-order evidence. Examples are statements about the amount and correctness of first-order evidence, the argumentative role of first-order evidence as well as the properties of the argumentative structure into which first-order evidence is embedded. In this thesis, analyses are performed for three examples of higher-order evidence, namely the inferential density of the dialectical structure at a certain time step, and, for every person, the amount of evidence claims and the ratio of true evidence claims.

Among today's epistemologists, there is a lively debate about the nature of higher-order evidence. It is characterized at least in two different ways: Higher-order

evidence is considered as evidence about (*i*) the character of the evidence or (*ii*) an agent's capacities for responding rationally to the evidence. According to Talbott (2016b), higher-order evidence is defeating evidence of a certain kind. It is a certain kind of undercutting defeater, namely a reliability defeater as described in (Pollock, 1984).

Second, I connect higher-order evidence in the sense of Talbott (2016b) with my own notion of higher-order evidence. Presuppose that participants of a debate form their beliefs according to confirmation. In this case, higher-order evidence in the sense of Talbott (2016b) is everything defeating the reliability of this cognitive process. Betz (2015) has shown that higher-order evidence, namely inferential density and the amount of evidence, allows us to estimate the reliability of absolute confirmation as a veritistic indicator for the truth of a hypothesis. However, these results are limited. First, Betz (2015) assumes a totally true body of evidence. Second, Betz (2015) only considers absolute confirmation as a veritistic indicator. The first part of my thesis expands upon (Betz, 2015) not only by considering (*i*) partly incorrect bodies of evidence but also (*ii*) relevance confirmation measures.

Among today's epistemologists, there is also a lively debate about the bearing of higher-order evidence. For example, is there a situation in which evidence of evidence (for some hypothesis h) is itself evidence for h? This thesis is claimed by some epistemologists, see for example (Feldman, 2005). However, it is also contested by some others, see for example (Fitelson, 2012). Presuppose that confirmation is another example of higher-order evidence and there is a situation in which confirmation reliably indicates the truth of a hypothesis. Then, there is a situation in which evidence of evidence (for some hypothesis h) is itself evidence for h.

1.1.3. Veritistic Formal Epistemology

Part one of my thesis, that is chapter 2, shows how higher-order evidence allows us to infer the reliability of a veritistic indicator, namely confirmation. In doing so, it contributes to the program of veritistic social epistemology as introduced in (Goldman, 2003). Veritistic social epistemology is concerned with how groups of persons can track down the truth, not at last due to interacting with one another. According to Goldman (2003), there are attitudes towards sentences which possess a fundamental veritistic value. For example, knowledge, error and ignorance have a fundamental veritistic value of 1, 0 and 0.5, respectively. Practices possess a veritistic

value insofar as they promote the acquisition of fundamental veritistic value, most of the time and for most persons. A veritistic indicator is used in a doxastic practice in order to indicate how to change a belief system. The reliability of a veritistic indicator is assessed via the veritistic value of the associated practice.

In this thesis, the reliability of a veritistic indicator is assessed via a statistical hypothesis test based on Monte-Carlo simulations. The reliability of a veritistic indicator improves, if significance decreases and power increases. The hypothesis $\neg h$ is tested against the alternative hypothesis h. For every statistical test, there is a critical region, that is a region where $\neg h$ has to be rejected. Here, it is chosen such that the power equals 0.25. Hence, this thesis expands upon (Betz, 2015) by assessing the reliability of confirmation in a new way.

Assessing the influence of higher-order evidence on the reliability of confirmation as a veritistic indicator questions not only pure reliabilism, but also bayesianism. Bayesian epistemologists claim that being justified in believing a hypothesis h depends both on the evidence as well as on h's prior probability, see for example (Talbott, 2016a). Pure reliabilist epistemologists claim that being justified in believing a hypothesis h only depends on the reliability of the cognitive process giving rise to the belief in h, see for example (Goldman and Beddor, 2021). Assessing the reliability of confirmation as a veritistic indicator comprises bayesian as well as reliabilist influences.

There are several re-evaluating principles, stating that I have to re-evaluate my former beliefs in light of certain evidence about the process producing these beliefs. Part one of my thesis underpins certain re-evaluating principles, as for example the $MERF$ principle as introduced by Talbott (2016b) and the integration principle as introduced by Christensen (2008), in a twofold way. First, it assesses the reliability of special cognitive processes, namely certain modes of belief formation according to confirmation. Second, it identifies reliability-relevant categorizations of these processes in terms of higher-order evidence. According to the integration principle, object-level beliefs must reflect meta-level beliefs about the reliability of the cognitive process. According to the $MERF$ principle, a belief must be revised, if it does not equal the reliability of the cognitive process (unless there is some categorization of this process such that it does equal the reliability of the cognitive process, at least approximately). According to Talbott (2016b), the reliability of a cognitive process is given by the expected relative frequency of truths among beliefs which

are produced by this very process.

1.1.4. Applied Epistemology as Rational Reconstruction

Part two of my thesis, that is chapter 3, reconstructs a historic scientific debate, namely the great Devonian controversy, using once again the theory of dialectical structures. In doing so, it reveals relations between evidence and hypotheses. There are several concepts of philosophy of science which my reconstruction of the great Devonian controversy possibly illustrates, for example the theory-ladenness of observational statements, Duhemian underdetermination as stated in (Duhem, 1954) or the the struggle about standardizing methodological rules for generating empirical statements.

The great Devonian controversy spans approximately from 1834 to 1841. Dating some strata in Devonshire is its start and end point. However, the great Devonian controversy is much more than a local debate. First, participants and observations come from all over Europe and North America. Second, its impacts are far-reaching, not only laterally but also temporally. As a result of the great Devonian controversy, several things have been established, namely (*i*) a new geological period, the Devon, (*ii*) a new dating method by means of characteristic fossil assemblages and (*ii*) the idea of a constant piecemeal change in fauna and flora.

During the great Devonian controversy, participants infer different hypotheses about the age of all the older strata in Devonshire from evidential beliefs often changing. These inferences are based on so-called mineralogical and fossil criteria, connecting the mineralogical character and fossil content of certain strata with certain geological ages, respectively. In the end, there is a consensus between the main participants, not only regarding the dating of all the older strata, but also most of the evidential statements including a certain criterion. How do participants change their beliefs such that they finally reach a consensus?

What is consensus? Are there different kinds? How does consensus change with time? Bramson et al. (2017) assess consensus respectively polarization in terms of groups and propose a variety of polarization measures, as for example community fragmentation, size parity, group consensus and distinctness. These measures answer very different questions: How many groups can be defined? How are participants distributed over groups? To what extent do positions of members of the same group

differ? Are there shared beliefs between members of different groups? For practical reasons, my analyses of the great Devonian controversy center on community fragmentation and size parity. However, using a similarity measure to define groups endogenously, group consensus as well as distinctness are assessed as well.

For every time step, not only groups are identified exogenously by accepting a certain piece of evidence, but endogenously using degrees of similarity between dating hypotheses respectively bodies of evidence. This way, my thesis enhances (Rudwick, 1988). In this thesis, are groups the same as those in (Rudwick, 1988)? Is this true for exogenous as well as endogenous clustering? One prominent point of my polarization analysis in terms of endogenously defined groups is the separation between dating hypotheses and bodies of evidence. This is highly motivated by striving for elucidating relations between hypotheses and bodies of evidence. Are similarity spectra of dating hypotheses and bodies of evidence similar? Do similarity dynamics of dating hypotheses and bodies of evidence coincide? Note also that similarity analyses may illustrate Kuhn (1983) stating that controversies are not only triggered, but also resolved by inter-paradigmatic exchange of arguments. Albeit, the illustration thesis relies on two assumptions, namely that (*i*) a dating hypothesis and a body of evidence constitute a paradigm and (*ii*) changes in similarity are caused by argumentation.

Is there some connection between individual belief change and confirmation? For the great Devonian controversy, this thesis quantifies the notion of confirmation in three different ways, namely using $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. In doing so, it reveals degrees of confirmations and confirmation dynamics. Are there differences between confirmation measures? Only between two of them or all of them? Both in value and relative changes? Does the great Devonian controversy start and end with all main participants accepting a dating hypothesis with a maximal degree of confirmation, independent of a certain confirmation measure?

How do participants of the great Devonian controversy change their beliefs individually? Answers to this question may pick up several concepts of philosophy of science, for example falsification of a hypothesis by some evidence in the sense of Popper (1935), separation of hard core and auxiliary assumptions as introduced in (Lakatos, 1970) or a model of piecemeal and reluctant belief change as developed in (Laudan, 1984).

Is rationality in belief change related with some kind of evidential support? In this

thesis, evidential support is spelled out in terms of confirmation, that is $DOJ(h|e)$, $Z_{DOJ}(h, e)$ and $F_{DOJ}(h, e)$. For the great Devonian controversy, do the following principles of rational belief change apply?

- ($RAT1$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it does not decrease the dating hypothesis's degree of evidential support.
- ($RAT2$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it maximizes the dating hypothesis's degree of evidential support.

Here, a person maximizes her dating hypothesis's evidential support, if she chooses a dating hypothesis such that, given her body of evidence, there is no better confirmed dating hypothesis. For the great Devonian controversy, are individual belief changes rational according to one of these two principles? Are there individual belief changes which are rational according to neither of them? If so, presuppose that participants of the great Devonian controversy are nevertheless rational. Then, it follows that there are exceptions to these principles. This would confirm Feyerabend (1976) stating that there are no strict rules for scientists.

For the great Devonian controversy, what is the final consensus all about? In short, the final consensus is all about the agreement on a certain dating hypothesis, namely the main part of the Culm, its black limestone and the Non-Culm being Coal Measures, Mountain Limestone and Old Red Sandstone in age, respectively. There is no total agreement on a certain body of evidence. However, the intersection of all bodies of evidence at the final time step is quite large. For example, there is a total agreement on a certain fossil criterion as well as a certain temporal order of all the older strata in Devon. In this thesis, approachment to the final consensus dating hypothesis is assessed via the similarity between a person's dating hypothesis at a certain time step and the final dating hypothesis, respectively. Regarding approachment to the final consensus body of evidence, things are a little bit more complicated. First, in my reconstruction of the debate, a body of evidence is as small as possible, that is, it does not include sentences which are implied by other evidential beliefs and the dialectical structure. Second, not all sentences of the final consensus are part of the dialectical structure at each time step. Therefore, I take it that the approachment to the final consensus body of evidence is only adequately assessed by the similarity between the deductive closure of a person's body of ev-

idence at a certain time and the deductive closure of the finally shared evidential beliefs restricted to those sentences which are part of the dialectical structure at that time.

For participants of the great Devonian controversy, do roads to the final consensus differ? Are there strong similarities? For all participants of the debate, are similarity dynamics not monotonous, that is, approaching alternates with distancing the final consensus? This would be in line with Betz (2013) investigating consensus-conduciveness of controversial debates by means of multi-agent simulations. A prominent point of my analysis of final consensus formation is the separation between dating hypotheses and bodies of evidence. This is a refinement of the analysis of final consensus formation given in (Rudwick, 1988). For all participants of the great Devonian controversy, does approaching the final consensus in terms of dating hypotheses not imply an approachment in terms of bodies of evidence, and vice versa?

For the great Devonian controversy, does an individual belief change decrease similarity with the final consensus, iff it is not rational? Here as before, the same two principles of rational belief change are considered, relating rationality with evidential support. Hence, this question translates into the two following ones: Does a belief change decrease similarity with the final consensus, iff it decreases evidential support? Does a belief change decrease similarity with the final consensus, iff it does not maximize evidential support?

Is maximizing evidential support well designed to approach the final consensus? Is this a rule with exceptions? If so, how can these exceptions be characterized? In terms of higher-order evidence? Remember that, in the first part of my thesis, analyses are performed for three examples of higher-order evidence, namely the inferential density of the dialectical structure at a certain time step, and, for every person, the amount of evidence claims and the ratio of true evidence claims. For the great Devonian controversy, the correctness of first-order evidence is assessed in terms of similarity with the final consensus body of evidence.

Is maximizing evidential support a reliable process for approaching the final consensus, that is, do most dating hypotheses maximizing evidential support have a sufficiently high degree of similarity with the final dating hypothesis? What influences the reliability of this process? Similarity with the finally shared evidential beliefs? Are there differences between confirmation measures? Are there differences

between different phases of the debate?

During early phases of the great Devonian controversy, maximizing evidential support is it well designed to approach the final consensus, independent of a certain person? Is the great Devonian controversy closed when, as a result of argumentation, evidence accumulation and belief changes, maximizing evidential support is maximally well designed in approaching the final dating hypothesis, independent of a certain person? My thesis answers these questions. In doing so, it may illustrate the compromise model as introduced in (Kitcher, 1993) in a new way. The compromise model concerns the closure of major scientific debates and embodies ideas of rationalism as well as anti-rationalism. In (Kitcher, 1993), the great Devonian controversy serves as an illustrative example for the compromise model. As Kitcher (1993) puts it, having read (Rudwick, 1988), it is clear that the first three conditions of the compromise model are met, but "issues around $C4$ and $C5$ are more murky". My analyses shed new light on these very issues.

1.2. Research Questions and Methods

1.2.1. Veritistically Valuable Reasoning with False Evidence

In the following, research questions of part one, that is chapter 2, are listed.

- $(V1)$ For a totally correct body of evidence, does higher-order evidence influence the relative frequency of truths among hypotheses with a certain degree of confirmation, only if using a relevance confirmation measure?
- $(V2)$ Independent of higher-order evidence, is absolute confirmation a more reliable indicator than relevance confirmation? Does higher-order evidence influence the reliability of confirmation as a veritistic indicator? Are there differences between absolute and relevance confirmation?
- $(V3)$ Are there situations where confirmation is a reliable veritistic indicator?
- $(V4)$ Are there situations where confirmation is no reliable veritistic indicator?

To answer these questions, 1000 simulated debates are drawn from (Betz, 2013) and analyzed in the following way in terms of confirmation, using $DOJ(h|e)$, $Z_{DOJ}(h,e)$ or $F_{DOJ}(h,e)$.

1. For every debate, inferential density is calculated, truth is set and an evidence stream is generated. As in (Betz, 2015), truth setting as well as evidence stream generation makes use of random numbers. An evidence stream is an accumulative list of lists of evidential beliefs. In contrast to (Betz, 2015), a certain ratio of these evidential beliefs are false.
2. For every hypothesis of a certain debate, the degree of confirmation is calculated given (*i*) a certain amount of evidential beliefs, (*ii*) a certain ratio of true evidential beliefs and (*iii*) a certain inferential density.
3. A a statistical hypothesis test is performed with confirmation as test statistic. The hypothesis $\neg h$ is tested against the alternative hypothesis h.
   a) The critical region is chosen such that the power of the test equals 0.25.
   b) The significance of the test is calculated.

Part one of this thesis, that is chapter 2, is organized as follows: A veritistic approach to argumentation is outlined in sec. 2.1.1, introducing the concept of a reliable veritistic indicator. The theory of dialectical structures, especially the concept of justification, is introduced in sec. 2.1.2. Why confirmation is considered a promising candidate for a reliable veritistic indicator is motivated in sec. 2.1.3. Three different notions of confirmation are presented in sec. 2.1.4. The concept of higher-order evidence is outlined in sec. 2.1.5. The methodology of the debate analyses is exposed in sec. 2.2.1 and results of these analyses are unfolded in sec. 2.2.2.

1.2.2. Historic Scientific Reasoning with False Evidence

In the following, research questions of part two, that is chapter 3, are listed.

- ($H1.1$) Does the the great Devonian controversy illustrate the concept of theory-ladenness?
- ($H1.2$) Does the great Devonian controversy illustrate Duhemian underdetermination as stated in (Duhem, 1954)?
- ($H1.3$) Does the great Devonian controversy illustrate the struggle about standardizing methodological rules for generating empirical statements?
- ($H1.4$) Does the great Devonian controversy start and end with all main participants accepting a dating hypothesis with a maximal degree of confirmation

(given a certain evidence), that is with a degree of 1, independent of a certain confirmation measure?

- ($H1.5$) For most time steps and participants, are $DOJ(h|e)$ and $Z_{DOJ}(h,e)$ rather similar, both in value and relative changes, and much smaller than 1?
- ($H1.6$) For most time steps and participants, are $DOJ(h|e)$ and $F_{DOJ}(h,e)$ rather unsimilar, both in value and relative changes?
- ($H2.1$) Are groups the same as those in (Rudwick, 1988)? Is this true for exogenous as well as endogenous clustering? How do results depend on the similarity threshold?
- ($H2.2$) Are similarity spectra of dating hypotheses and bodies of evidence similar? Do similarity dynamics of dating hypotheses and bodies of evidence coincide?
- ($H2.3$) Is the average degree of similarity maximal at the final step, not at last due to argumentation?
- ($H3.1$) Given that participants of the great Devonian controversy are rational, do their belief changes dis-confirm strict falsificationism in the sense of Popper (1935)?
- ($H3.2$) For the great Devonian controversy, are there beliefs which can be considered as hard-core assumptions in the sense of Lakatos (1970)?
- ($H3.3$) Does the great Devonian controversy illustrate Laudan (1984) stating that beliefs are not revised as a whole, but rather in a piecemeal and reluctant way?
- ($H4.1$) Do participants of the great Devonian controversy are rational according to some principles of rational belief relating rationality with evidential support? Are there differences between confirmation measures?
- ($H4.2$) Do shared belief changes and changes in the dialectical structure more often decrease and less often maximize evidential support than individual belief changes? Are there differences between confirmation measures?
- ($H5.1$) Does approaching alternate with distancing the final consensus?
- ($H5.2$) Does approaching the final consensus in terms of dating hypotheses not imply an approachment in terms of bodies of evidence, and vice versa?

- Is maximizing evidential support well designed to approach the final consensus, that is, do most dating hypotheses maximizing evidential support have a sufficiently high degree of similarity with the final dating hypothesis? ($H5.3$) Are there big differences between persons? ($H5.4$) Are there differences between confirmation measures?
- ($H5.5$) During early phases of the great Devonian controversy, is maximizing evidential support well designed to approach the final consensus, independent of a certain person?
- ($H5.6$) Is the great Devonian controversy closed when, as a result of argumentation, evidence accumulation and belief changes, maximizing evidential support is maximally well designed in approaching the final dating hypothesis, independent of a certain person?

To answer these questions, time span of the great Devonian controversy is discretized and the following is performed for every time step:

1. The dialectical structure is reconstructed, identifying auxiliary assumptions, inferential relations, argument types and clusters.
2. For every main participant and the final consensus, a dating hypothesis and evidential beliefs are identified.
3. Groups of persons are defined exogenously as well as endogenously using a newly introduced similarity measure.
4. For every main participant, her dating hypothesis' degree of confirmation given all her evidential beliefs is calculated.
5. For every main participant, similarity with the final consensus is assessed using the previously introduced similarity measure.
6. For every main participant, the relative frequency of dating hypotheses which are sufficiently similar with the final dating hypothesis among all those dating hypotheses maximizing evidential support is determined.

Part two of my thesis, that is chapter 3, is organized as follows: The great Devonian controversy is shortly introduced and characteristics of its reconstruction are outlined in sec. 3.1, referring to common concepts in philosophy of science as theory-ladenness of observational data, Duhemean underdetermination, mini- and maxi-theories as well as newly introduced concepts as atomic dating hypotheses,

shared background beliefs, bodies of evidence and two kinds of time slicing. Next, in sec. 3.2, polarization and polarization dynamics are assessed in terms of groups, which are defined endogenously or exogenously. Further, individual belief changes are contrasted with some prominent philosophical views on this topic. For every main participant, sec. 3.3 presents confirmation dynamics. Results are compared for three different notions of confirmation and belief changes analyzed in terms of confirmation. Finally, in sec. 3.4, for every main participant, roads to the final consensus are analyzed. Special focus lies on relations between similarity with the final consensus and confirmation as well as the reliability of a certain kind of forming beliefs according to confirmation and its reliability defeaters.

2. Veritistically Valuable Reasoning with False Evidence

How can we reason with false evidence and reliably infer hypotheses? In the formal framework employed in this section, this question translates into: Given some hypothesis and a body of evidence which is partly incorrect, is there a reliable veritistic indicator, i.e. something that most of the time correctly indicates the truth or falsity of a hypothesis?

To answer this question a Monte-Carlo analysis is performed on simulated debates drawn from (Betz, 2013). These debates instantiate a formal model of argumentation, namely the theory of dialectical structures as introduced in (Betz, 2010). In doing so, this analysis contributes to the program of computational philosophy as for example presented in (Grim and Singer, 2020).

The analysis probes, given a body of evidence with a certain ratio of false evidence claims, the reliability of confirmation as a veritistic indicator. Assessing the reliability of confirmation as a veritistic indicator comprises bayesian as well as reliabilist influences. Based on theoretical claims about confirmation as partial entailment taken from Crupi and Tentori (2013) and Crupi and Tentori (2014), the veritistic merit of three different confirmation measures is analyzed, namely $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. The first one measures absolute confirmation and the two latter ones measure relevance confirmation.

Here, the reliability of a veritistic indicator is assessed via a statistical hypothesis test which is performed in accordance with James (2006). The reliability of a veritistic indicator improves, if the significance decreases or the power increases. The hypothesis $\neg h$ is tested against the alternative hypothesis h based on Monte-Carlo simulations. For every statistical test, there is a critical region, that is a region where $\neg h$ has to be rejected which is chosen in two different ways, namely such that the power equals 0.25 or the significance equals 0.05.

It is in the scope of the analysis to investigate influences of higher-order evidence on the reliability of confirmation as a veritistic indicator. What is higher order evidence? First some words on my own notion of higher-order evidence. I consider a body of evidence as first-order evidence and a statement about first-order evidence as higher-order evidence. Examples are statements about the amount and correctness of first-order evidence, the argumentative role of first-order evidence as well as the properties of the argumentative structure into which first-order evidence is embedded. In this thesis, analyses are performed for three examples of higher-order evidence, namely the inferential density of the dialectical structure at a certain time step, and, for every person, the amount of evidence claims and the ratio of true evidence claims.

This analysis shows how higher-order evidence allows us to infer the reliability of a veritistic indicator. In doing so, it contributes to the program of veritistic social epistemology as introduced in (Goldman, 2003). Assessing the influence of higher-order evidence on the reliability of confirmation as a veritistic indicator questions not only pure reliabilism, but also bayesianism.

The main findings of the analysis are:

- ($V1$) The relative frequency of truths among hypotheses with a certain degree of confirmation. It shows that, for a totally correct body of evidence, higher-order evidence influences the relative frequency, only if using a relevance confirmation measure.
- ($V2$) The reliability of confirmation as a veritistic indicator. Independent of higher-order evidence, it shows that absolute confirmation is a more reliable indicator than relevance confirmation. Higher-order evidence influences the reliability of confirmation as a veritistic indicator and there are differences between absolute and relevance confirmation.
  - ($V2.0$) As the ratio of false evidence claims increases, the reliability decreases.
  - ($V2.1$) As the inferential density increases, the reliability increases. For a body of evidence including false evidence claims and absolute confirmation, the truth of $V2.1$ depends on the amount of evidence.
  - ($V2.2$) As more and more evidence is accumulated, the reliability increases. For a body of evidence including false evidence claims, the truth

of $V2.2$ depends on the inferential density. However, there are exceptions. First, for a body of evidence including false evidence claims and a certain kind of relevance confirmation, $V2.2$ holds, independent of the inferential density. Second, for a sufficiently large amount of false evidence claims and absolute confirmation, $V2.2$ does not hold, independent of the inferential density.

  - ($V2.3$) As more and more evidence is accumulated, differences between absolute and relevance confirmation decrease.
  - ($V2.4$) As the inferential density increases, differences between absolute and relevance confirmation decrease.

- ($V3$) Situations in which its degree of confirmation is evidence for the confirmed hypothesis. There are situations where confirmation is a reliable veritistic indicator. It shows that absolute confirmation is more often a reliable veritistic indicator than relevance confirmation. For a sufficiently small amount of true evidence claims, confirmation is no reliable veritistic indicator, independent of a certain confirmation measure.
- ($V4$) Exceptions to the epistemic rule of forming beliefs according to confirmation. There are situations where it is not rational to form beliefs according to confirmation, namely those where confirmation is no reliable indicator for the truth of a hypothesis.

Some remarks on $V1$. Presuppose that subjects respond to some hypothesis in accordance with its degree of confirmation given some body of evidence. In what circumstances is this process reliable? For a totally true body of evidence, Betz (2015) has shown that the inferential density as well as the amount of evidence allow us to estimate the reliability of absolute confirmation as a veritistic indicator for the truth of a hypothesis. Betz (2015) assesses reliability in terms of the relative frequency of truths among hypotheses with a certain degree of confirmation. Hence, $V1$ expands upon (Betz, 2015) by considering (i) partly incorrect bodies of evidence and (ii) relevance confirmation measures. Further, there are several re-evaluating principles, stating that one has to re-evaluate one's former beliefs in light of certain evidence about the process producing these beliefs. $V1$ underpins a certain re-evaluating principle, namely the $MERF$ principle, as introduced by Talbott (2016b), in a twofold way. First, it assesses the reliability of a special cognitive process, namely belief formation according to confirmation. Second, it identifies reliability-

relevant categorizations of this very process in terms of higher-order evidence.

Some remarks on $V2$. $V2$ expands upon (Betz, 2015) not only by considering (i) partly incorrect bodies of evidence and (ii) relevance confirmation measures but also by assessing the reliability of confirmation as a veritistic indicator via significance and power of a corresponding statistic hypothesis test. $V2$ underpins a certain re-evaluating principle, namely the integration principle as introduced by Christensen (2008), in a twofold way. First, it assesses the reliability of a special cognitive process, namely belief formation according to a statistic hypothesis test with confirmation as test statistic. Second, it identifies reliability-relevant categorizations of this very process in terms of higher-order evidence.

Some remarks on $V3$. Let us suppose that confirmation is evidence of evidence. Then, $V3$ shows that there are situations, in which evidence of evidence (for some hypothesis h) is itself evidence for h, with a situation being characterized by some other examples of higher-order evidence, namely the inferential density of the dialectical structure, the amount of evidence and the ration of true evidence claims. This thesis is claimed by some epistemologists, see for example (Feldman, 2005). However, it is also contested by some others, see for example (Fitelson, 2012).

Some remarks on $V4$. For all three confirmation measures, there are situations where confirmation is no reliable veritistic indicator, with a situation being characterized in terms of higher-order evidence, namely the inferential density of the dialectical structure, the amount of evidence and the ration of true evidence claims. Therefore, adjusting one's belief according to confirmation is not always rational. Presume that adjusting one's beliefs according to confirmation is an epistemic rule. Then, my analysis shows that there are exceptions to this rule. Hence, it confirms Feyerabend (1976) stating that there is no scientific rule without any exceptions. However, it does not support relativism in the sense of "anything goes".

This section is organized as follows: A veritistic approach to argumentation is outlined in sec. 2.1.1, introducing the concept of a reliable veritistic indicator. The theory of dialectical structures, especially the concept of justification, is introduced in sec. 2.1.2. Why confirmation is considered a promising candidate for a reliable veritistic indicator is motivated in sec. 2.1.3. Three different notions of confirmation are presented in sec. 2.1.4. The concept of higher-order evidence is outlined in sec. 2.1.5. The methodology of debate analyses is exposed in sec. 2.2.1 and results of these analyses are unfolded in sec. 2.2.2.

2.1. Propaedeutics

2.1.1. Veritistic Indicators

This section introduces some key concepts of veritism. These are taken from Goldman (2003) who elaborates veritistic social epistemology. Veritistic social epistemology is concerned with how groups of persons can track down the truth, not at last due to interacting with one another. Goldman (2003) focuses on certain modes of interaction and forms of communication as for example argumentation, testimony, computer-mediated and scholarly communication.

According to Goldman (2003), there are attitudes towards sentences which possess a fundamental veritistic value. For example, knowledge, error and ignorance have a fundamental veritistic value of 1, 0 and 0.5, respectively. Practices possess a veritistic value insofar as they promote the acquisition of fundamental veritistic value, most of the time and for most persons.

Given a body of evidence which is partly incorrect, is there a veritistically valuable practice of inferring hypotheses? As a first example, consider a person obeying the following norm.

> If e and $I(e, h)$, then h.

Here, e is some body of evidence, h is some hypothesis, and I is some relation between e and h. Obeying this norm is a veritistically valuable practice, if the relation is a reliable veritistic indicator for the truth of h. Something is a veritistic indicator, if it is used in a doxastic practice in order to indicate how to change a belief system. The reliability of a veritistic indicator is assessed via the veritistic value of the associated practice. The veritistic value of the corresponding practice increases with the reliability of the veritistic indicator.

Given a body of evidence which is partly incorrect, is there a reliable veritistic indicator for the truth of a hypothesis? As a second example, consider a person obeying the following norm.

> If e and $I(e, h) \in \omega$, then h.

Here, e is some body of evidence, h is some hypothesis, I is some function with arguments e and h, and ω is some interval. Here, I take it that the veritistic value of this norm respectively the reliability of I as a veritistic indicator can be measured in

two different ways. First, the veritistic value of this norm decreases with increasing probability of false hypotheses falling into the interval. Second, the veritistic value of this norm increases with increasing probability of true hypotheses falling into the same interval. Note that, in the realm of statistic hypothesis testing, these probabilities are known as significance and power of a statistic hypothesis test, see definition 2.1.1 and for example James (2006).

Definition 2.1.1: Significance and Power a Statistic Hypothesis Test

The significance and power of a statistic hypothesis test with test statistic $I(h, e)$, critical region ω and null hypothesis $\neg h$ are defined as follows.

$$\alpha = P(I(h, e) \in \omega | \neg h)$$
$$(1 - \beta) = P(I(h, e) \in \omega | h)$$

Here, P is some probability function

Remember that, following the rules of statistic hypothesis testing, one has to reject the null hypothesis, if the test statistic falls into the critical region. Depending on each other, significance and power cannot be calculated both at the same time. Choosing one of them determines the critical region. Here, I take it that a power of 0.25 is sufficiently large and a significance of 0.05 is sufficiently small to ensure the reliability of I as a veritistic indicator.

Given a body of evidence which is partly incorrect, is there some veritistic indicator such that the significance of a corresponding hypothesis test is sufficiently small given a power of 0.25? At least in case of a totally correct body of evidence, justification is a reliable veritistic indicator as long as the amount of evidence as well as inferential density are sufficiently large. More information on this can be found in Betz (2015) or in the following chapters.

2.1.2. Dialectical Structures and Justification

Here as well as in (Betz, 2015), justification is understood in accordance with the theory of dialectical structures, that is a formal model of complex argumentation. In the following, some key concepts of the theory of dialectical structures are shortly introduced. For more information on this theory see (Betz, 2010).

According to the theory of dialectical structures, every state of a debate can be represented by a dialectical structure and the positions of its proponents. A *dialectical structure* consists of sentences and deductive valid arguments possibly attacking and supporting one another. Let A and B be two deductively valid arguments and s and t be two sentences. Then, it holds:

- s is contradictory to t, iff s and t can neither be both true nor false at the same time.
- s is contrary to t, iff s and t cannot be both true at the same time.
- s attacks B, iff s is contradictory to one of B's premises.
- s supports B, iff s is one of B's premises.
- A attacks s, iff A's conclusion is contradictory to s.
- A supports s, iff A's conclusion is s.
- A attacks B, iff A's conclusion is contradictory to one of B's premises.
- A supports B, iff A's conclusion is one of B's premises.

A *position* assigns truth-values to sentences. If truth-values are assigned to all sentences in a dialectical structure, then the position is called *complete*, otherwise *partial*. A complete position is *dialectically consistent,* iff it (i) assigns complementary truth values to a sentence and its negation and (ii) considers conclusions of arguments with true premises as true. The *inferential density* of a dialectical structure increases with decreasing number of all complete and dialectical consistent positions, compare definition 2.1.2.

The *absolute degree of justification* of some sentence depends on the number of all complete and dialectical consistent positions as well as the number of all complete and dialectically positions extending this very sentence, compare definition 2.1.3. There is also a *conditional degree of justification* of some sentence given another sentence or a set of sentences. In contrast to the absolute degree of confirmation, it does not depend on the number of all complete and dialectical consistent positions, but on the number of all complete and dialectically consistent positions extending both the sentence which is justified and the set of sentences which justifies, compare definition 2.1.3. According to Betz (2012), it holds that degrees of justification

(DOJ) satisfy the probability axioms of Kolmogorov.[1] Therefore, it holds :

$$\begin{aligned} DOJ(\neg h) &= 1 - DOJ(h) \\ DOJ(\neg h|e) &= 1 - DOJ(h|e) \end{aligned} \tag{2.1}$$

$$DOJ(h|e) = \frac{DOJ(e|h)DOJ(h)}{DOJ(e)} \tag{2.2}$$

Definition 2.1.2: Inferential Density

Let n and σ be the number of all sentences and all complete and dialectically consistent positions within a dialectical structure τ. Then, the inferential density, $D(\tau)$, is defined as follows.

$$D(\tau) = \frac{n - \lg(\sigma)}{n}$$

Definition 2.1.3: Degrees of Justification

Let h and e be partial positions within a dialectical structure τ. Let σ be the number of complete and dialectical consistent positions in τ and σ_h the number of complete and dialectical positions extending h. Then, the degree of justification of h, $DOJ(h)$, and the conditional degree of justification of h given e, $DOJ(h|e)$, are defined as follows.

$$\begin{aligned} DOJ_\tau(h) &= \tfrac{\sigma_h}{\sigma} \\ DOJ_\tau(h|e) &= \tfrac{\sigma_{h\&e}}{\sigma_e} \end{aligned}$$

2.1.3. Reasoning with True Evidence

This section summarizes (Betz, 2015) which analyses the reliability of justification as a veritistic indicator. As in this thesis, justification is understood in accordance with the theory of dialectical structures. The veritistic analyses are performed over 1000 simulated debates drawn from (Betz, 2013).[2] Every simulated debate contains 20

[1] Basic axioms and theorems of probability can be found for example in (Schurz, 2006).

[2] There is a video illustrating (Betz, 2013). It is accessible via https://www.youtube.com/watch?v=aIFq8McAoZY.

propositions and consists of a series of dialectical structures. Two adjacent dialectical structures differ only in one argument, compare example 2.1.1.

In (Betz, 2013), simulations are multi-agent simulations. According to Winsberg (2019), this type of simulation is most common in the social and behavioral sciences. Other simulation types are for example equation based simulations and Monte-Carlo simulations. The first one is common in physical sciences and aims to solve differential equations approximately. The second one comprises all those algorithms where the approximation of a mathematical quantity, such as for example an integral or a statistical parameter, makes use of random numbers. Note that Monte Carlo simulations are often thought of either calculating a mathematical quantity or imitating a physical system, as for example a debate with 6 proponents and 20 sentences. However, as stated in (Beisbart and Norton, 2012), the second case reduces to the first one. Using random numbers, Monte Carlo simulations generate a subset of all possible trajectories of the system. Based on this subset, statistical parameters can be calculated which are approximations of the true statistical parameters, as for example the mean conditional degree of justification of true hypotheses.

(Betz, 2013) simulates a debate with 6 proponents and 20 sentences as follows. In the beginning of a debate, there is no argument, but there are six proponents holding complete positions which are randomly chosen. In the course of a debate, proponents introduce arguments according to some argumentation mechanism. In (Betz, 2013), there are five different argumentation mechanisms. There is a most simple one positing that new arguments are devised randomly. Further, there are four mechanisms assuming that arguments are introduced such that a certain argumentation rule is obeyed. A proponent responds to a new dialectical structure according to some update mechanism specifying the evolution of her position.

Results shown in (Betz, 2015) are based on the assumption that arguments are devised randomly. However, there is also a robustness analysis assuring that results do not hinge on the type of argumentation mechanism. (Betz, 2015) uses (Betz, 2013) merely as a source of dialectical structures with a certain inferential density. Proponents only figure in as much as bodies of evidence can in principle be associated with them.

In order to perform veritistic analyses, for every debate, a certain complete position has to be chosen as the truth. In (Betz, 2015), for every debate, the truth is randomly chosen among all complete and dialectically consistent positions at the

final time step, compare example 2.1.2. To investigate the veritistic merit of evidence accumulation, for every dialectical structure and hypothesis, an evidence stream is generated, that is a body of evidence which accumulates. Here, evidence is a presupposed sentence symbolized by some number $1, ..., 20$. In (Betz, 2015), in order to generate an evidence stream with regard to some hypothesis h, two steps are performed consecutively. First, a maximal body of evidence is generated, that is a set of 19 sentences not including h. The order of sentences is chosen randomly. Second, a sequence of 20 subsets is generated such that for two adjacent subsets X and $X+1$, it holds: (i) the cardinality of $X+1$ is $k+1$, with k being the cardinality of X and (ii) the intersection of X and $X+1$ is X. The first subset of the evidence stream is the empty set, the last one is the maximal body of evidence with respect to h. So, as evidence accumulates, more and more evidence claims of this maximal body of evidence are taken under consideration, compare Algorithmus 1 and example 2.1.3.

Due to argument construction, truth setting and evidence accumulation using random numbers, the simulation in (Betz, 2015) is a Monte-Carlo one. For different sizes of the body of evidence as well as inferential densities, not only the mean conditional degree of justification of true hypotheses is calculated, but also the relative frequency of true hypotheses among all hypotheses with a certain degree of conditional justification.

Betz (2015) shows that evidence accumulation and argumentation improves the reliability of justification as a veritistic indicator. However, these results are limited. First, Betz (2015) assumes a totally correct body of evidence. Yet, in real debates, the body of evidence often contains false evidence claims. Second, Betz (2015) only considers justification as a veritistic indicator. Yet, in real debates, especially in scientific ones, beliefs are often formed according to some notion of relevance confirmation. Both limitations will be removed in this chapter. Note also that Betz (2015) does not assess the reliability of a veritistic indicator via significance and power of a corresponding statistical hypothesis test.

Example 2.1.1: Adjacent Dialectical Structures

Here, two adjacent dialecticals stuctures, τ_1 and τ_2, are shown.

$$\begin{aligned}\tau_1 = \quad & (12 \wedge 19 \implies 10) \wedge (\neg 18 \wedge \neg 10 \implies 20) \wedge (\neg 13 \wedge \neg 8 \implies 6) \wedge \\ & (\neg 10 \wedge \neg 1 \implies \neg 11) \wedge (\neg 19 \wedge 15 \implies \neg 16) \wedge (5 \wedge 6 \implies 17) \wedge \\ & (\neg 5 \wedge 13 \implies 1) \wedge (\neg 15 \wedge \neg 6 \implies 7) \wedge (\neg 12 \wedge 15 \implies 2) \wedge \\ & (\neg 17 \wedge 18 \implies 7) \\ \tau_2 = \quad & \tau_1 \wedge (\neg 8 \wedge \neg 7 \implies \neg 5)\end{aligned}$$

Example 2.1.2: Possible Truth Candidate

Here, a possible truth is shown, that is a complete position which is dialectically consistent all of the time.

$$\begin{aligned}t = \quad & \neg 1 \wedge \neg 2 \wedge \neg 3 \wedge \neg 4 \wedge \neg 5 \wedge 6 \wedge 7 \wedge 8 \wedge 9 \wedge 10 \wedge \neg 11 \wedge \neg 12 \\ & \wedge \neg 13 \wedge 14 \wedge \neg 15 \wedge 16 \wedge 17 \wedge \neg 18 \wedge \neg 19 \wedge \neg 20\end{aligned}$$

Algorithmus 1 : Evidence Accumulation

Input : The truth T
Output : Evidence accumualtion with respect to hypothesis h

```
1 def efunc(T,h):
2     Generate a pseudo-random sample;
3     Remove h;
4     Generate subsets such that their elements accumulate;
```

Example 2.1.3: Evidence Accumulation

Here, accumulation of a body of evidence is performed with respect to hypothesis 9:
(i) Random sample of the truth with hypothesis 9 removed

$$\neg 13 \wedge \neg 1 \wedge 17 \wedge \neg 19 \wedge \neg 4 \wedge \neg 8 \wedge 14 \wedge 5 \wedge 10 \wedge \neg 15 \wedge \neg 2$$
$$\wedge \neg 11 \wedge \neg 18 \wedge \neg 12 \wedge \neg 3 \wedge 7 \wedge 6 \wedge \neg 20 \wedge 16$$

(ii) Generation of subsets such that their elements accumulate

$$\{\{\}, \{\neg 13\}, \{\neg 13 \wedge \neg 19\}, \{\neg 13 \wedge \neg 19 \wedge 16\}, ..., \{\neg 13 \wedge \neg 19 \wedge 16$$
$$\wedge 5 \wedge 14 \wedge \neg 20 \wedge 7 \wedge \neg 4 \wedge 17 \wedge \neg 2 \wedge 6 \wedge \neg 15 \wedge 10 \wedge \neg 11$$
$$\wedge \neg 12 \wedge \neg 8 \wedge \neg 1 \wedge \neg 18 \wedge \neg 3\}\}$$

2.1.4. Confirmation

Satisfying the Kolmogorov' axioms, degrees of justification can be used to spell out Bayesian confirmation theories. Hence, confirmation can be understood in terms of justification. A similar approach can be found in (Festa, 1999) where confirmation is understood in terms of an epistemic probability, namely *plausibility*.

Bayesian confirmation theory is a probabilistic theory of confirmation. The following insights into this theory are taken from Crupi (2021). The basic postulate of probabilistic confirmation is called formality and states that there is a function which depends on nothing but the probability distribution over the algebra generated by the hypothesis and the evidence. There are two prominent special cases of this function called absolute confirmation and relevance confirmation, see definitions 2.1.4 and 2.1.5.

Definition 2.1.4: Absolute Confirmation

Be P some probability function, h some hypothesis, e some evidence. Then, the degree of absolute confirmation is given by
$CONF_P(h, e) = g[P(h \mid e)]$
The more $P(h \mid e)$ is larger (smaller) than 0.5, the more h is confirmed (disconfirmed) by e with respect to P.

Definition 2.1.5: Relevance Confirmation

Be P some probability function, h some hypothesis, e some evidence. Then the degree of relevance confirmation is given by
$CONF_P(h, e) = g[P(h \mid e), P(h)]$
The more $P(h \mid e)$ is larger (smaller) than $P(h)$, the more h is confirmed (dis-confirmed) by e with respect to P.

There are other theories of confirmation such as Hempelian confirmation and hypothetico-deductivism. According to the first theory, a hypothesis is confirmed by some evidence, iff the evidence entails a suitable instantiation of the hypothesis. For example, a white (non-white) swan confirms (dis-confirms) the hypothesis that all swans are white. According to the second theory, a hypothesis is confirmed (dis-confirmed) by some evidence, iff the hypothesis entails the evidence (the negation of the evidence) with the help of suitable auxiliary hypotheses and assumptions. Being qualitative theories of confirmation, both theories are not suited for my analyses. For more information on these two theories of confirmation see for example (Hempel, 1945a), (Hempel, 1945b), or (Huber, 2008).

To choose some confirmation measure out of the multitude of discussed Bayesian confirmation measures, the following condition is imposed: There have to be good reasons to consider confirmation as an extension of the concept of deductive entailment. Here, I take it that good reasons are those given by Crupi and Tentori (2013) and Crupi et al. (2007). According to them, it holds that, if some confirmation measure can be considered as an extension of the concept of deductive entailment, then two conditions are met. These two conditions are shortly presented in definition 2.1.6. According to Crupi et al. (2007), among a variety of relevance confirmation measures, only $F_P(h, e)$ and $Z_P(h, e)$ meet the first condition, and only $Z_P(h, e)$ meets both conditions. Note that the absolute confirmation measure $P(h|e)$ meets both conditions, too.

Definition 2.1.6: Confirmation as Partial Entailment

Be $CONF_P(h,e)$ some confirmation measure, (e,h) and $(e,\neg h)$ deductive arguments and ν some function assigning all $(e,h)/(e,\neg h)$ the same value $x/-x$ with $x>0$. Then, it holds:
(Ex_1) $CONF_P(h,e)$ assigns $(e,h)/(e,\neg h)$ always a higher/lower value than some inductive argument.
(Ex_2) $CONF_P(h,e)$ mirrors a symmetry μ, iff ν mirrors μ, that is $\nu(e,h)CONF_P(e,h) = \nu(\mu(e,h))CONF_P(\mu(e,h))$
Here, a symmetry $\mu(e,h)$ is a function negating one or both arguments or changing their positions.

In the following, Bayesian confirmation theory is spelled out using degrees of justification as a probability function, that is

$$CONF_P(h,e) \rightarrow CONF_{DOJ}(h,e)$$

Hence, in the following, Bayesian confirmation measures are $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. In accordance with Crupi and Tentori (2010), Branden Fitelson (2004) and Kemeny and Oppenheim (1952), the two latter ones are defined as follows:[3]

Definition 2.1.7: $F_{DOJ}(h,e)$

Be DOJ degrees of justfication, h some hypothesis and e some evidence.

$$F_{DOJ}(h,e) \equiv \begin{cases} L_{DOJ}(h,e), & 0 < DOJ(h|e) < 1 \\ 1, & DOJ(h|e) = 1 \wedge DOJ(e) \neq 0 \\ -1, & DOJ(h|e) = 0 \end{cases}$$

with

$$L_{DOJ}(h,e) = \frac{DOJ(e|h) - DOJ(e|\neg h)}{DOJ(e|h) + DOJ(e|\neg h)}$$

[3]Note that, in this thesis and in contrast to its first definition, $Z_{DOJ}(h,e)$ accounts for non-contingent hypotheses, too.

Definition 2.1.8: $Z_{DOJ}(h,e)$

Be DOJ degrees of justfication, h some hypothesis and e some evidence.

$$Z_{DOJ}(h,e) = \begin{cases} \frac{DOJ(h|e)-DOJ(h)}{1-DOJ(h)}, & DOJ(h|e) \geq DOJ(h) \\ \frac{DOJ(h|e)-DOJ(h)}{DOJ(h)}, & DOJ(h|e) < DOJ(h) \\ 1, & DOJ(h) = 1 \\ -1, & DOJ(h) = 0 \end{cases}$$

2.1.5. Higher Order Evidence

What is evidence? In philosophy of science, there are several different answers. For example, a piece of evidence is considered to be a stimulus of a sensory receptor, a current mental state or a thing that is known, compare (Quine, 1969), (Feldman and Conee, 1985) or (Williamson, 2000). In this thesis, a piece of evidence is a sentence which is presupposed. The totality of all pieces of evidence is called the body of evidence. Using the theory of dialectical structures, a hypothesis' degree of justification depends on the body of evidence. Hence, evidence makes a difference to what one is justified in believing, as stated in (Kelly, 2016).

What one is justified in believing, is it entirely determined by one's evidence? Once more, in philosophy of science, there are several different answers. There are some philosophers answering this question in the affirmative, see for example (Feldman and Conee, 1985). However, there are also others. Bayesian epistemologists claim that being justified in believing a hypothesis h depends both on the evidence as well as on h's prior probability, see for example (Talbott, 2016a). Pure reliabilist epistemologists claim that being justified in believing a hypothesis h only depends on the reliability of the cognitive process giving rise to the belief in h, see for example (Goldman and Beddor, 2021). Assessing the reliability of confirmation as a veritistic indicator comprises bayesian as well as reliabilist influences. Assessing the influence of higher-order evidence on this reliability questions not only pure reliabilism, but also bayesianism.

What is higher-order evidence? Among today's epistemologists, there is a lively debate about the nature of higher-order evidence. It is characterized at least in two different ways: Higher-order evidence is considered as evidence about (i) the char-

acter of the evidence or (*ii*) an agent's capacities for responding rationally to the evidence. My own notion of higher-order evidence is of the first kind. I consider a body of evidence as first-order evidence and a statement about first-order evidence as higher-order evidence. Examples are statements about the amount and correctness of first-order evidence, the argumentative role of first-order evidence as well as the properties of the argumentative structure into which first-order evidence is embedded. In this thesis, analyses are performed for three examples of higher-order evidence, namely the inferential density of the dialectical structure at a certain time step ($D(\tau)$), and, for every person, the amount of evidence claims ($|e|$) and the ratio of true evidence claims ($|e_T|/e$).

Is higher-order evidence evidence of a peculiar kind? There are some philosophers answering this question in the affirmative. According to Christensen (2010), "respecting [higher-order evidence] can apparently force a person to fall short in certain ways, by having beliefs that fail to respect logic or basic inductive support relations". The same view he finds in (Field, 2005) where "[h]igher-order evidence sometimes required agents to violate or compromise certain rational ideals".

Is higher-order evidence defeating evidence of a certain kind? According to Talbott (2016b), it is a certain kind of undercutting defeater, namely a reliability defeater in the sense of Pollock (1984). He defines a reliability defeater (RD) via the following reason scheme:

> "(RS) For any x, $P(x)$ is a *prima facie* reason for $Q(x)$.
> (RD) I am in circumstances of type C, and something's being P in circumstances of type C is not a reliable indication that it is Q."

Note that according to Pollock (1984), a reliability defeater can be defeated as well, namely by believing RDD, which is defined as follows.

> (RDD) I am in circumstances of type C^* narrower than C, and something's being P in circumstances of type C^* is a reliable indication that it is Q."

Hence, according to (Talbott, 2016b, p. 3121), higher-order evidence "is evidence of the unreliability of the causal processes responsible for the undercut belief".

There is a connection between this very notion of higher-order evidence and mine. Presuppose that participants of a debate form their beliefs according to confirmation. According to (Talbott, 2016b, p. 3121), higher-order evidence is everything,

which defeats the reliability of this cognitive process. Betz (2015) has shown that the inferential density of a dialectical structure as well as the amount of evidence allow us to estimate the reliability of absolute confirmation as a veritistic indicator for the truth of a hypothesis. However, these results are limited. First, Betz (2015) assumes a totally true body of evidence. Second, Betz (2015) only considers absolute confirmation as a veritistic indicator. Third, Betz (2015) assesses reliability in terms of the relative frequency of truths among hypotheses with a certain degree of confirmation.

What about a body of evidence including false evidence claims and relevance confirmation? For a relevance confirmation measure and a body of evidence including false evidence claims, what are the relative frequencies of truths among hypotheses with certain degrees of confirmation? And how does higher-order evidence influence these frequencies? The next chapter provides answers to these questions.

In this thesis, the reliability of confirmation as a veritistic indicator is assessed via significance and power of a corresponding statistical hypothesis test, compare the previous sub-chapter. For a body of evidence and a hypothesis test with confirmation as the test statistic, what is the significance of the test given a power of 0.25? And how does higher-order evidence influence this significance? Are there differences between absolute and relevance confirmation? The next chapter provides answers to these questions, too.

Among today's epistemologists, there is also a lively debate about the bearing of higher-order evidence. For example, is there a situation in which evidence of evidence (for some hypothesis h) is itself evidence for h? This thesis is claimed by some epistemologists, see for example (Feldman, 2005). However, it is also contested by some others, see for example (Fitelson, 2012). Presuppose that confirmation is another example of higher-order evidence and there is a situation in which confirmation is a reliable veritistic indicator, that is the significance is sufficiently small. Then, there is a situation in which evidence of evidence (for some hypothesis h) is itself evidence for h.

Reasoning with higher-order evidence, what for? There are several re-evaluating principles, stating that I have to re-evaluate my former beliefs in light of certain evidence about the process producing these beliefs. Consider the following examples:

- **Higher-Order Defeat** by Lasonen-Aarnio (2014): Evidence that a cognitive process producing a doxastic state S as output is flawed has a defeating force

with respect to S.

- **Integration** by Christensen (2008): An agent's object-level beliefs must reflect the agent's meta-level beliefs about the reliability of the cognitive processes underlying her object-level beliefs.
- **MERF** by Talbott (2016b): An agent's confidence assignment of c in P is in disequilibrium if: There is some CP_1 such that $ERF(c, CP_1) \not\approx c$ and there is no CP_2 with $CP_2 \leq CP_1$ such that $ERF(c, CP_2) \approx c$. Here, CP_x is a reliability-relevant categorization of the causal processes responsible for c and $ERF(c, CP_x)$ is the expected relative frequency of truths among propositions assigned confidence of c on the basis of a causal process of kind CP_x.

Today, many epistemologists agree on the use of a re-evaluating principle. However, there are also exceptions, see for example (Lasonen-Aarnio, 2014). According to Lasonen-Aarnio (2014), there only exists a coherent notion of epistemic rationality, if there is an epistemic rule with no exceptions or re-evaluating principles have to be rejected, at least sometimes.

Example 2.1.4: Argument with Epistemic Modesty

(EM) There is no epistemic rule without any exceptions.
(A1) A re-evaluating principle is an epistemic rule.
(CER) There is a coherent notion of epistemic rationality.

(REP) There are situations where re-evaluating principles can rationally be rejected.

Notice that according to Talbott (2016b), the $MERF$ principle itself implies that, given our evidence, our degree of confidence in it should be less than 0.5. What follows from rejecting a re-evaluating principle? Rejecting the $MERF$ principle, one no longer endorses external calibration and accuracy maximization as epistemic goals of rational degrees of confidence.

The first part of this thesis underpins re-evaluating principles in a twofold way. First, it assesses the reliability of a special cognitive process, namely belief formation according to some notion of confirmation. Second, it identifies reliability-relevant categorizations of this very process in terms of higher-order evidence.

2.2. Confirmation as a Veritistic Indicator

Given a body of evidence, which is partly incorrect, is confirmation a reliable veritistic indicator and does higher-order evidence allow us to estimate its reliability? As a part of my thesis, this question is answered based on veritistic analyses of simulated debates. The set up of these analyses is shortly described in sec. 2.2.1. Their results are presented in sec. 2.2.2.

2.2.1. Setting

A Monte-Carlo analysis is performed on simulated debates drawn from (Betz, 2013). In doing so, this analysis contributes to the program of computational philosophy. According to Grim and Singer (2020), computational philosophy comprises all philosophical research making use of computational techniques. As Grim and Singer (2020) puts it: "The idea is simply to apply advances in computer technology and techniques to advance discovery, exploration and argument within any philosophical area."

Veritistic analyses are performed over 1000 simulated debates drawn from (Betz, 2013). These debates instantiate a formal model of argumentation, namely the theory of dialectical structures as introduced in (Betz, 2010) and shortly described in sec. 2.1.2. As in (Betz, 2015), results are based on the assumption that arguments are devised randomly. My thesis uses (Betz, 2013) merely as a source of dialectical structures with a certain inferential density. Proponents only figure in as much as bodies of evidence can in principle be associated with them.

For every debate, truth is set and an evidence stream is generated. As in (Betz, 2015), truth setting as well as evidence stream generation makes use of random numbers. An evidence stream is an accumulative list of lists of evidential beliefs. In order to study reasoning with false evidence, and in contrast to Betz (2015), a certain ratio of these evidential beliefs are false. See sec. 2.1.3 and Algorithmus 2 for a detailed listing of the algorithms of truth setting and evidence accumulation. As argument construction, truth setting and evidence accumulation use random numbers, my simulations are Monte-Carlo ones.

As in (Betz, 2015), simulations calculate approximately statistical parameters. However, these are not the same. In this thesis, for every hypothesis of a certain debate,

the degree of confirmation is calculated given a certain amount of evidential beliefs, ratio of true evidential beliefs and inferential density. Results can be visualized as a histogram regarding degrees of confirmation, see sec. A.1.

A statistical test is performed with confirmation as test statistic. The hypothesis to be tested is $\neg h$ and it is tested against the alternative hypothesis h. The test statistic is a confirmation measure, namely $DOJ(h|e)$, $Z_{DOJ}(h,e)$ or $F_{DOJ}(h,e)$. First, the critical region is chosen such that the power (significance) equals 0.25 (0.05). Second, the significance (power) is calculated. If the significance (power) is sufficiently small (large), then there are sufficiently few false (many true) hypotheses within the critical region.[4] Hence, in this case, a degree of confirmation within the critical region is a reliable veritistic indicator for the truth of a hypothesis.

Based on a subset of all possible trajectories of the modeled system, namely a debate with 6 proponents, 20 sentences and false evidence, the significance and power as calculated in my thesis only approximate the true significance and power of a statistical test with confirmation as test statistic. Here, I take it that this approximation is good enough, that is, that the sample is large enough. Even given that the sample is large enough, is the chosen model a good enough representation of a real debate with false evidence? Certainly, the model used in this thesis reflects a compromise between the best description of the phenomena and computational tractability. Nevertheless, the second part of my thesis strives to offer some kind of validation by reconstructing and analyzing a historic debate with false evidence in terms of confirmation and consensus, see chapter 3.

Algorithmus 2 : Evidence Accumulation - False Evidence

Input : The truth T
Output : Evidence Accumualtion with Respect to Hypotheses h

```
1 def efunc2(T,h):
2     For a certain number of sentences in T, reverse truth-value assignments;
3     Generate a pseudo-random sample;
4     Remove h;
5     Generate subsets such that their elements accumulate;
```

[4] Remember that a power of 0.25 and a significance of 0.05 correspond to 25 percent of all true hypotheses and 5 percent of all false hypotheses having a degree of confirmation within the critical region, respectively, compare also sec. 2.1.1.

2.2.2. Results

In this chapter, results are presented which provide answers to the following questions:

1. What is the relative frequency of truths among hypotheses with a certain degree of confirmation? Does higher-order evidence influence these relative frequencies?
2. Is confirmation a reliable veritistic indicator? For a body of evidence and a hypothesis test with confirmation as test statistic, what is the significance, given a power of 0.25? Does higher-order evidence influence the reliability of confirmation as a veritistic indicator? Does higher-order evidence influence the significance of the test?[5]
3. Are there situations in which its degree of confirmation is evidence for the confirmed hypothesis? Are there situations in which the significance of a corresponding hypothesis test with confirmation as tests statistics is sufficiently small?

Here, analyses take into account three examples of higher-order evidence, namely the amount of first-order evidence, the ratio of true first-order evidence claims, and the inferential density of the dialectical structure in which first-order evidence is embedded. Confirmation is spelled out in three different ways, using an absolute confirmation measure and two relevance confirmation measures, namely $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. Is there a difference between absolute and relevance confirmation?

First, some words on results answering the first question. For the three discussed confirmation measures, Fig. 2.1, Fig. 2.2 and Fig. 2.3 show grids of plots with rows and columns differing in inferential density ($D(\tau)$) and amount of evidence ($|e|$). Every single point shows the relative frequency of truths among hypotheses within a certain interval of confirmation. This is done for three different ratios of true evidence claims, namely 0.6 (blue), 0.8 (green) and 1.0 (violet).

Violet points in Fig. 2.1 reproduce results of (Betz, 2015). For a completely correct body of evidence, the relative frequency of truths among hypotheses with a certain

[5] The following questions are answered in sec. A.2: For a body of evidence and a hypothesis test with confirmation as test statistic, what is the power of the test given a significance of 0.05? Does higher-order evidence influence the power of the test?

degree of confirmation virtually equals this very degree of confirmation, regardless of the amount of evidence and the inferential density. As green and blue points show, this is only true for totally correct bodies of evidence.

As the ratio of false evidence claims increases, there are more and more true hypotheses with a low degree of absolute confirmation as well as false hypotheses with a high one. As a consequence, for not totally correct bodies of evidence and absolute confirmation, the relative frequency of truths among hypotheses with a certain degree of confirmation no longer equals this very degree. The corresponding difference increases with confirmation getting extremal. Note that for not totally correct bodies of evidence and absolute confirmation, the relative frequency of truths among hypotheses with a certain degree of confirmation is still a linear function of this very degree. The only exception are extremal values of confirmation which figure as discontinuities. Note also that for not totally correct bodies of evidence and absolute confirmation, amount of evidence and inferential density do have some influence on the relative frequency of truths among hypotheses with a certain degree of confirmation, albeit only a minor one. As amount of evidence respectively inferential density increases, the relative frequency of truths among hypotheses with a certain degree of absolute confirmation increases.

Comparing Fig. 2.1, Fig. 2.2 and Fig. 2.3, it shows that there are some similarities between absolute and relevance confirmation. As amount of evidence respectively inferential density increases, the difference between the relative frequency of truths among hypotheses with a certain degree of confirmation and this very degree increases. This is true for all three confirmation measures, however not to the same extent. Using $F_{DOJ}(h, e)$, this difference increases most drastically.

Comparing Fig. 2.1, Fig. 2.2 and Fig. 2.3, it shows that there are considerable differences between absolute and relevance confirmation. For a totally correct body of evidence, the relative frequency of truths among hypotheses with a certain degree of relevance confirmation equals this very degree only very approximately or not at all. Further, for a totally correct body of evidence, inferential density and the amount of evidence have some influence on the relative frequencies of truths among hypotheses with a certain degree of relevance confirmation. Independent of the ratio of true evidence claims, the relative frequency of truths among hypotheses with a certain degree of relevance confirmation is no linear function of this very degree. However, most of the time, this function is approximately continuous. As for absolute confir-

mation and not totally correct bodies of evidence, extremal values of confirmation figure as uncontinuities.

Second, some words on results answering the second question. Fig. 2.4 shows a grid of plots with rows and columns differing in confirmation measures and ratio of true evidence claims, respectively. Every single plot shows a grid with rows and columns differing in inferential density and amount of evidence. The significance of a hypothesis test with confirmation as test statistic and a power of 0.25 is represented using a color function. Turning from blue over yellow to red, the color corresponds to values between 0 and 0.24.

Independent of higher-order evidence, it shows that the significance is smallest using $DOJ(h|e)$ and largest using $F_{DOJ}(h, e)$. For absolute confirmation, the color spectrum ranges from a light shade of blue to a dark one. For relevance confirmation, the color spectrum comprises shades of yellow, too. The darkest shade of yellow is much darker using $F_{DOJ}(h, e)$ than $Z_{DOJ}(h, e)$. Hence, $DOJ(h|e)$ is a more reliable indicator for the truth of a hypothesis than $Z_{DOJ}(h, e)$, and $Z_{DOJ}(h, e)$ is a more reliable indicator for the truth of a hypothesis than $F_{DOJ}(h, e)$.

Higher-order evidence has some influence on the significance of the test. Consider the following claims:

- ($V2.0$) As the ratio of false evidence claims increases, the reliability decreases.
- ($V2.1$) As the inferential density increases, the reliability increases.
- ($V2.2$) As more and more evidence is accumulated, the reliability increases.
- ($V2.3$) As more and more evidence is accumulated, differences between absolute and relevance confirmation decrease.
- ($V2.4$) As the inferential density increases, differences between absolute and relevance confirmation decrease.

According to Fig. 2.4, $V2.0$ holds for all confirmation measures and values of inferential density and amount of evidence. What about $V2.3$ and $V2.4$? As the inferential density respectively amount of evidence increases, differences between color spectra decrease. This is true for all pairs of confirmation measures. It is also true for all ratios of true evidence claims, however not to the same extent.

For a totally correct body of evidence, $V2.2$ and $V2.3$ hold independent of a certain confirmation measure, compare the first column of Fig. 2.4. As the inferential density

respectively amount of evidence increases, the significance decreases, that is the corresponding color turns more and more into a dark shade of blue. This is true for all three confirmation measures. Hence, argumentation and evidence accumulation improves the reliability of absolute as well as relevance confirmation as a veritistic indicator. Note that these results confirm those of Betz (2015), using a different way of assessing the reliability of a veritistic indicator.

As the ratio of true evidence claims decreases, things get a little bit more complicated. Compare the second and third column of Fig. 2.4. Only for relevance confirmation, $V2.1$ holds independent of the amount of evidence. For absolute confirmation and a sufficiently large amount of evidence, the significance of the test does not decrease with increasing inferential density. Instead of turning more and more into a dark blue, the color stays the same shade of blue respectively turns into a darker shade of yellow. Only using $F_{DOJ}(h, e)$, $V2.2$ holds independent of the inferential density. Using $Z_{DOJ}(h, e)$, the significance does not decrease with increasing amount of evidence, if the inferential density is sufficiently large. Instead of turning more and more into a dark blue, the color turns into a darker shade of blue respectively yellow. The same is true for absolute confirmation. Further, for absolute confirmation and a sufficiently small amount of true evidence claims, it holds that the significance does not decrease with increasing amount of evidence, independent of the inferential density.

Third, some words on results answering the third question. Fig. 2.5 shows a grid of plots with rows and columns differing in confirmation measures and ratio of true evidence claims, respectively. Every single plot shows a grid with rows and columns differing in inferential density and amount of evidence. The significance of a hypothesis test with confirmation as test statistic and a power of 0.25 is represented using a color function. Orange refers to a significance which is less or equals 0.05, and blue refers to one which is greater than that very value.

For all three confirmation measures, it shows that there are situations in which the significance of the test is less or equal 0.05. Here, a situation is characterized by higher-order evidence, namely the inferential density, the amount of evidence and the ratio of true evidence claims. Using $DOJ(h|e)$ and $F_{DOJ}(h, e)$, there is the largest and smallest amount of orange squares, respectively. Using $DOJ(h|e)$, $Z_{DOJ}(h, e)$ and $F_{DOJ}(h, e)$, there are eight, five and three orange squares, respectively. For a sufficiently small amount of true evidence claims, there are no orange squares at

all, independent of a certain confirmation measure. Using $F_{DOJ}(h, e)$, there are orange squares, only if the body of evidence is totally correct. Using $DOJ(h|e)$ or $Z_{DOJ}(h, e)$, there are orange squares for a ratio of true evidence claims of 0.8 and an inferential density of 0.45. For a totally correct body of evidence and a sufficiently high amount of evidence, there are orange squares independent of the inferential density. However, this is only true for absolute confirmation. For a totally correct body of evidence and a sufficiently small amount of evidence, there are no orange squares at all, independent of the inferential density. However, this is ´only true for relevance confirmation.

For all three confirmation measures, there are situations in which the significance is larger than 0.05, namely those represented by blue squares. Hence, it shows that there are situations where confirmation is no reliable veritistic indicator, with a situation being characterized by higher-order evidence, namely the inferential density, the amount of evidence and the ration of true evidence claims.

Consider the following epistemic rule defining rationality in terms of a hypothesis test as known from the previous subsection.

> ($RAT0$) Given some evidential beliefs, accepting a hypothesis is rational, if its degree of confirmation falls into the critical region of a statistic hypothesis test with confirmation as test statistic and a power of 0.25.

My analyses show that, most of the time, being rational in the sense of this epistemic rule fosters approaching the truth. However, there are situations, namely those with a significance unequal 0, where an agent being rational in this sense possibly accepts a false hypothesis. In these cases, being rational in this sense does not foster approaching the truth. This result relates to Feyerabend (1976) stating that there is no epistemic rule without any exceptions. He argues for this thesis in several ways. For example, he refers to scientific progress, claiming that, for every epistemic rule, there are some situations where violating it fosters scientific progress, compare (Feyerabend, 1976, p. 35).[6]

[6]Here, I take it that, given its context, “Fortschritt” can be translated and interpreted as “epistemic progress”.

Example 2.2.1: Argument with Scientific Progress

(FS1) Scientific progress has to be fostered.
(FS2) For all possible epistemic rules: There are some situations where violating the epsitemic rule fosters scientific progress.
—
(EM) For all possible epistemic rules: There are some situations where the epistemic rule has to be violated.

The recent results relate to $FS2$ by providing conforming instances. Nevertheless, this thesis must not support relativism in the sense of "anything goes".[7] Consider the following examples:

- There is an epistemic rule, $R1$, such that, for most situations, not violating $R1$ fosters scientific progress.
- There is an epistemic rule, $R2$, such that, for some situations, violating $R2$ fosters scientific progress, and all of these situations are known.
- There is an epistemic rule, $R3$, such that, for most of the situations studied so far, not violating $R3$ fosters scientific progress.

All these rules - $R1$, $R2$ and $R3$ - are rules with exceptions. However, they should not be dismissed at once. An agent aiming to foster scientific progress should not violate $R1$ in the long run. Further, he should not violate $R2$, except being in a situation, which constitutes a known exception. This is far from "anything goes", but nevertheless in accordance with Feyerabend (1976). Things are a little bit more difficult with $R3$. It is not sure that not violating $R3$ fosters scientific progress in the long run. Further, an agent studying a new situation can not exclude that violating $R3$ fosters scientific progress. What should he do? According to (Feyerabend, 1976, p. 45), he should regard $R3$, but carefully study the situation and show a general willingness to consider it as an exception.

The second part of this thesis, that is chapter 3, shows that real epistemic agents do not always form their beliefs according to confirmation. As will be shown, in some

[7] Feyerabend (1976) does not talk of *"relativism"*, but *"liberalism"*, namely liberal philosophies and principles, see for example page 252 and 307. Here, I take it that, given its context, "liberalism" refers to what is generally known among philosophers as "relativism". Feyerabend (1976) himself sometimes changes terminology, see for example remarks on *"anarchism"* and *"dadaism"* on page 33.

situations, there are non-epistemic reasons. However, as shown in this chapter, in some situations, there are epistemic reasons, too.

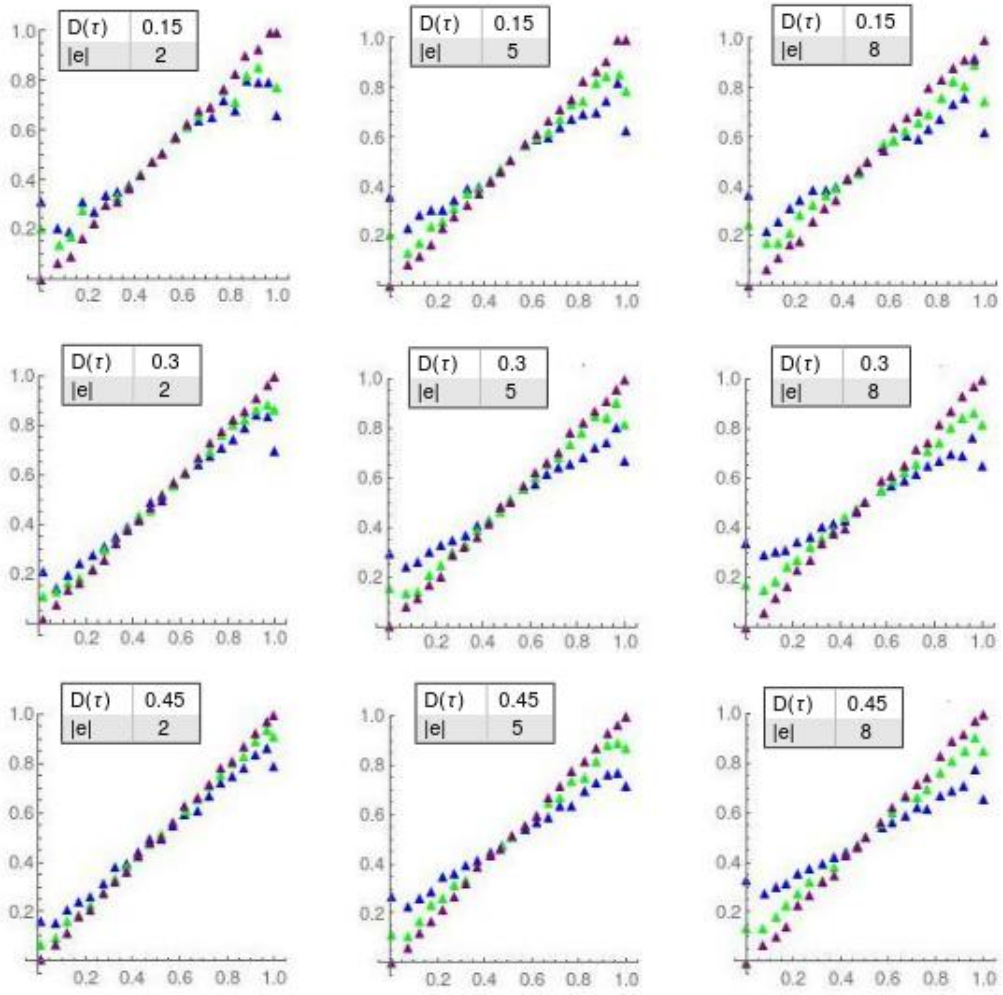

Figure 2.1.: Frequencies of Truths. A grid of plots is shown with rows and columns differing in inferential density ($D(\tau)$) and amount of evidence ($|e|$). Every single point shows the relative frequency of truths among hypotheses within a certain interval of $DOJ(h|e)$. This is done for three different ratios of true evidence claims, namely 0.6 (blue), 0.8 (green) and 1.0 (violet)

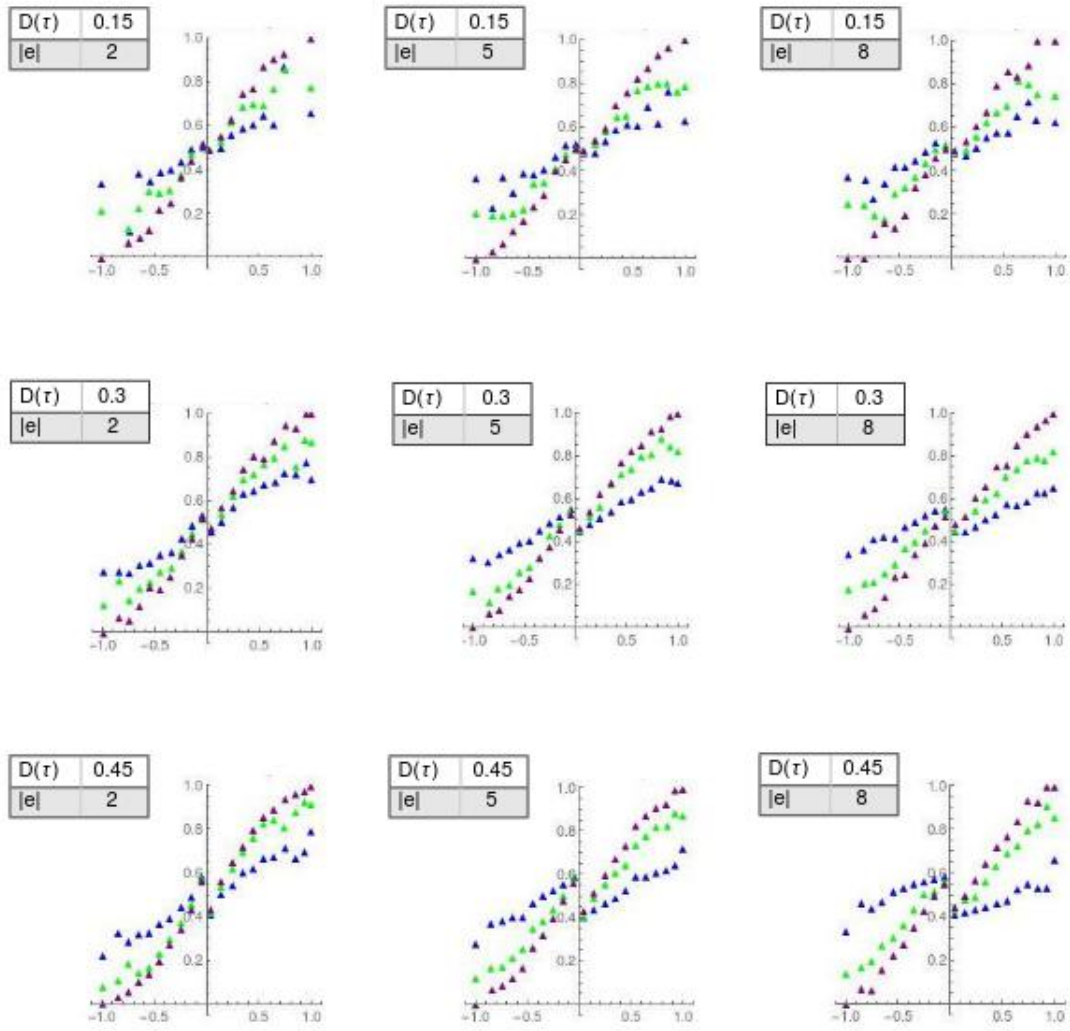

Figure 2.2.: Frequencies of Truths. A grid of plots is shown with rows and columns differing in inferential density ($D(\tau)$) and amount of evidence ($|e|$). Every single point shows the relative frequency of truths among hypotheses within a certain interval of $Z_{DOJ}(h, e)$. This is done for three different ratios of true evidence claims, namely 0.6 (blue), 0.8 (green) and 1.0 (violet)

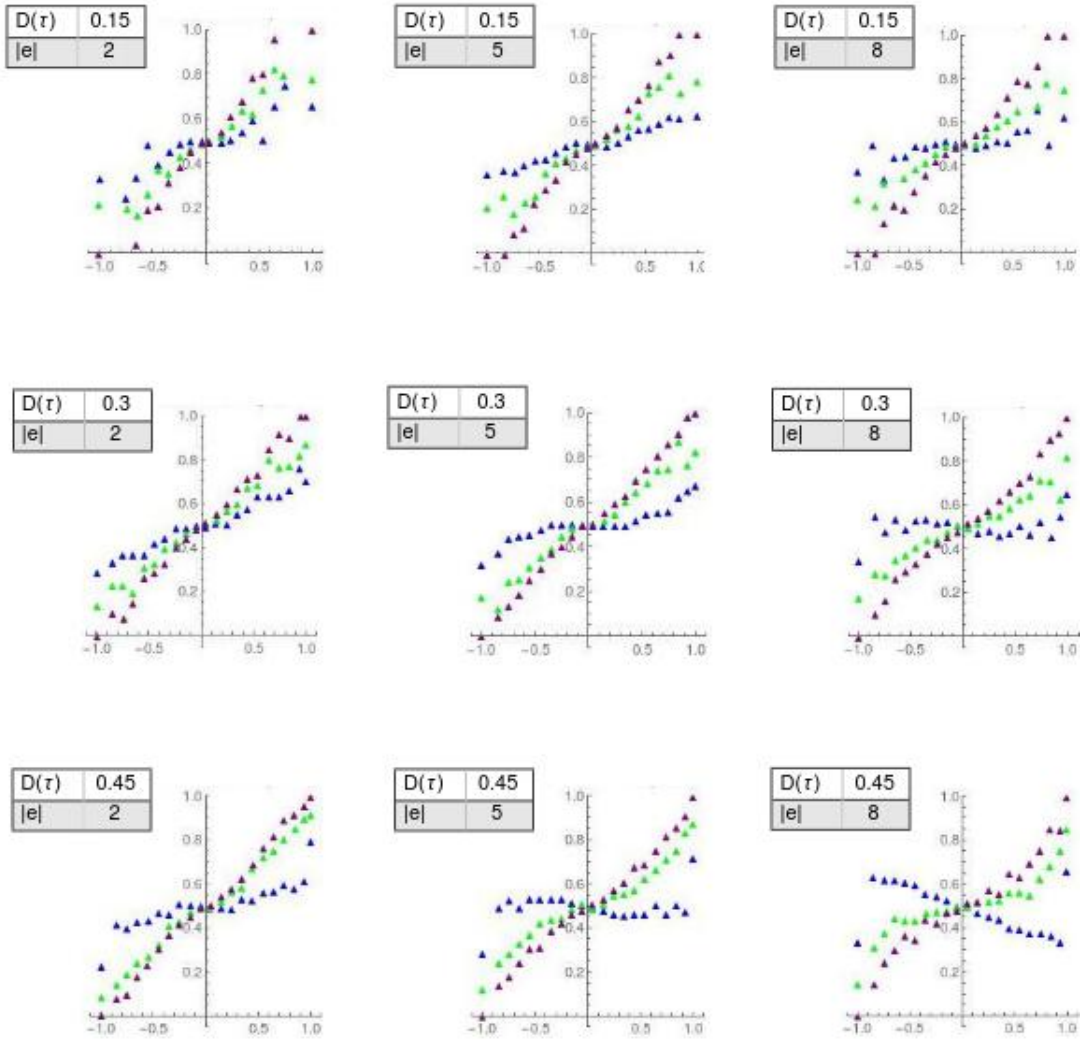

Figure 2.3.: Frequencies of Truths. A grid of plots is shown with rows and columns differing in inferential density ($D(\tau)$) and amount of evidence ($|e|$). Every single point shows the relative frequency of truths among hypotheses within a certain interval of $F_{DOJ}(h, e)$. This is done for three different ratios of true evidence claims, namely 0.6 (blue), 0.8 (green) and 1.0 (violet)

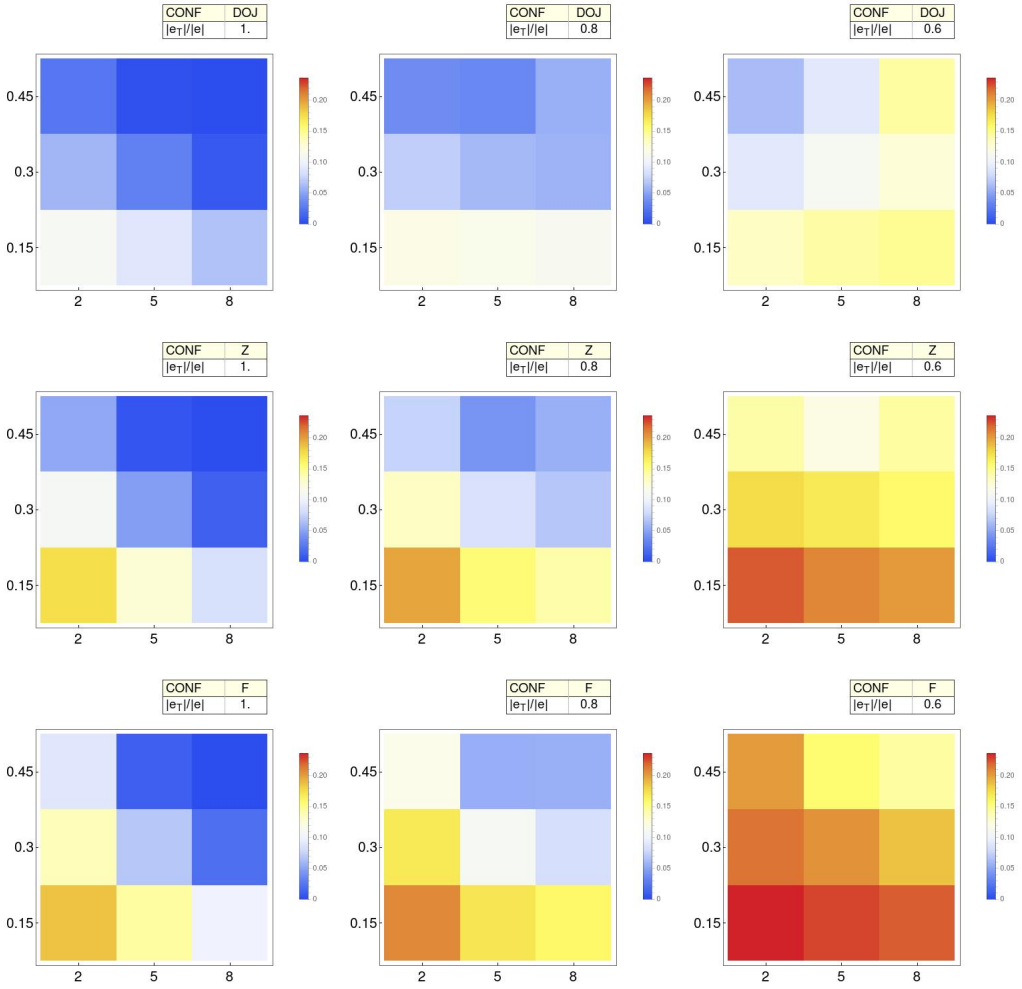

Figure 2.4.: Significance. A grid of plots is shown with rows and columns differing in confirmation measures and ratio of true evidence claims, respectively. From top to bottom, rows correspond to $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. From left to right, the ratio of true evidence claims decreases. Every single plot shows a grid with rows and columns differing in inferential density and amount of evidence. For every single plot, from bottom to top, the inferential increases and, from left to right, the amount of evidence increases. Significance is represented using a color function. Turning from blue over yellow to red, it increases.

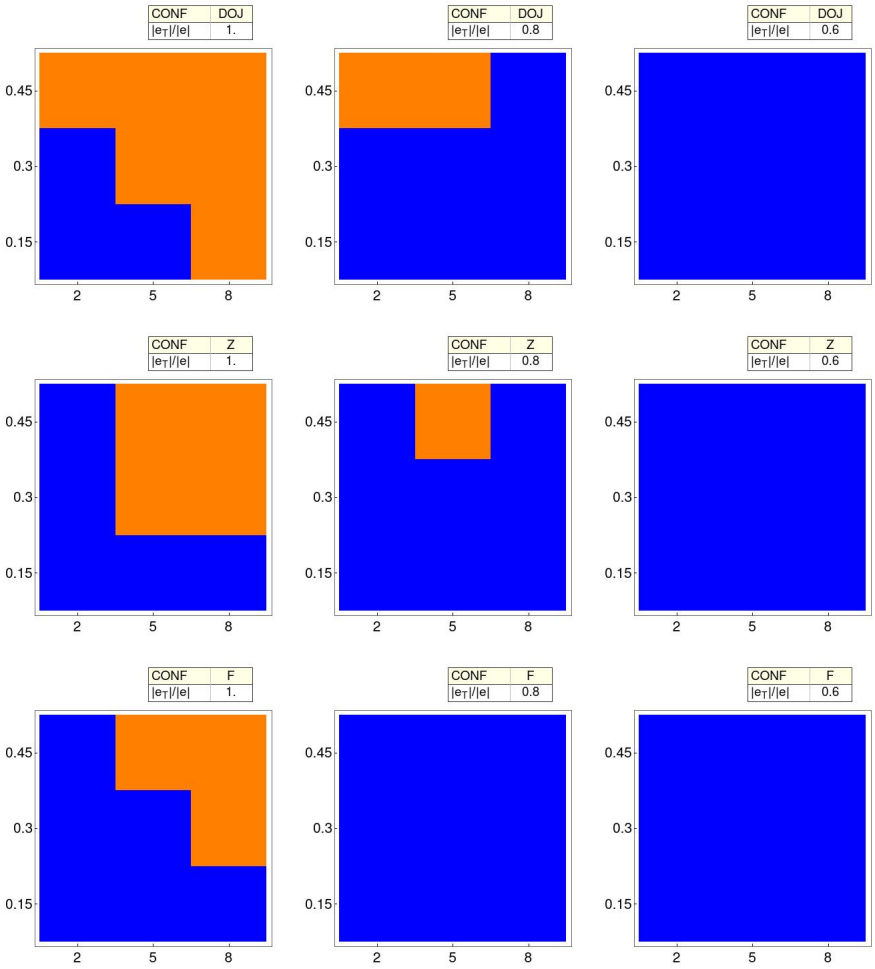

Figure 2.5.: Significance less or equal 0.05. A grid of plots is shown with rows and columns differing in confirmation measures and ratio of true evidence claims, respectively. From top to bottom, rows correspond to $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. From left to right, the ratio of true evidence claims decreases. Every single plot shows a grid with rows and columns differing in inferential density and amount of evidence. For every single plot, from bottom to top, the inferential increases and, from left to right, the amount of evidence increases. Orange refers to a significance which is less or equals 0.05.

3. Historic Scientific Reasoning with False Evidence

As an example of a valuable way of reasoning with false evidence, I chose the great Devonian controversy (GDC). Not at last due to its short analyses in (Kitcher, 1993). This is an exceptionally well documented scientific debate among 19th century geologists about the dating of all the older strata in Devonshire.

The debate spans approximately from 1834 to 1841. Dating some strata in Devonshire is its start and end point. However, the great Devonian controversy is much more than a local debate. First, participants and observations come from all over Europe and North America. Second, its impacts are far-reaching, not only laterally but also temporally. As a result of the debate, several things have been established, namely (*i*) a new geological period, the Devon, (*ii*) a new dating method by means of characteristic fossil assemblages and (*ii*) the idea of a constant piecemeal change in fauna and flora.

During the debate, participants infer different hypotheses about the age of all the older strata in Devonshire from evidential beliefs often changing. These inferences are based on so-called mineralogical and fossil criteria connecting the mineralogical character and fossil content of certain strata with certain geological ages, respectively. In the end, there is a consensus between the main participants, not only regarding the dating of all the older strata but also most of the evidential statements, including a certain fossil criterion. How do the main participants change their beliefs such that they finally reach a consensus? Is there some connection between individual belief change and confirmation? To answer these questions, time span of the debate is discretized and the following is performed for every time step:

1. The dialectical structure is reconstructed identifying auxiliary assumptions, inferential relations, argument types and clusters.

2. For every main participant and the final consensus, all evidential beliefs and a dating hypothesis are identified.
3. Groups of persons are defined exogenously as well as endogenously using a newly introduced similarity measure.
4. For every main participant, her dating hypothesis' degree of confirmation given all her evidential beliefs is calculated.
5. For every main participant, similarity with the final consensus is assessed using the previously introduced similarity measure.

These analyses foster understanding and deepen knowledge of several important philosophical concepts and issues:

- ($H1$) Relations between evidence and hypotheses
- ($H2$) Consensus and consensus dynamics
- ($H3$) Individual belief changes
- ($H4$) Rational belief change
- ($H5$) Consensus formation

Some words on $H1$, that is relations between evidence and hypotheses. Reconstructing the great Devonian controversy, the following shows:

- For all empirical statements of the great Devonian controversy, it holds: There is a dependence between its theoretical context and rational acceptance. Hence, the debate illustrates nicely the concept of theory-ladenness, which is uncontroversial in today's philosophy of science, see for example (Boyd and Bogen, 2021).
- ($H1.2$) For the great Devonian controversy, it shows that an empirical statement is implied by some mineralogical or fossil criterion, only if it is conjoined with some auxiliary assumptions. Therefore, the debate illustrates Duhemian underdetermination, compare (Duhem, 1954).
- ($H1.3$) There are several criteria and most of them are highly controversial most of the time. Only at the end, there is a criterion which all participants agree upon. Therefore, the great Devonian controversy illustrates the struggle about standardizing methodological rules for generating empirical statements.

Not only inferential relations between evidence and dating hypotheses are revealed, but also the notion of confirmation is quantified. Three different confirmation measures are compared, namely $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. See sec. 2.1.4 for their definition. Calculating degrees of confirmation, the following shows.

- ($H1.4$) The great Devonian controversy starts and ends with all main participants accepting a dating hypothesis with a maximal degree of confirmation (given a certain evidence), that is with a degree of 1, independent of a certain confirmation measure.
- ($H1.5$) For most time steps and participants, $DOJ(h|e)$ and $Z_{DOJ}(h,e)$ are rather similar, both in value and relative changes, and much smaller than 1.
- ($H1.6$) For most time steps and participants, $DOJ(h|e)$ and $F_{DOJ}(h,e)$ are rather unsimilar, both in value and relative changes. For most time steps and participants, $F_{DOJ}(h,e)$ increases with increasing $\frac{DOJ(h|e)}{DOJ(h)}$ and is fairly approximated by 1.

Some words on $H2$, that is consensus and consensus dynamics. For every time step, not only groups are identified exogenously by accepting a certain piece of evidence, but endogenously using degrees of similarity between dating hypotheses respectively bodies of evidence. This way, my thesis enhances (Rudwick, 1988). The main findings of these analyses are:

- ($H2.1$) For exogenous clustering, groups are never the same as those in (Rudwick, 1988). This is not true for endogenous clustering. Clustering maximally similar dating hypotheses, groups are the same as those in (Rudwick, 1988).
- ($H2.2$) Similarity spectra of dating hypotheses and bodies of evidence are quite similar, namely $[0.60, 1.00]$ and $[0.54, 0.99]$. However, similarity dynamics of dating hypotheses and bodies of evidence do not coincide:
  - Except during the middle section, there are always some persons accepting the same dating hypothesis. Never, not even at the end, there are two persons accepting the same body of evidence.
  - Several times, there are two persons accepting, at the same time, remarkably similar bodies of evidence but unsimilar dating hypotheses, and vice versa.
- ($H2.3$) The average degree of similarity is maximal at the final step, not at last due to argumentation. In so far as dating hypotheses and bodies of ev-

idence together constitute a paradigm, this result may be understood as an illustration of Kuhn (1983) stating that controversies are not only triggered, but also resolved by inter-paradigmatic exchange of arguments.

Some words on $H3$, that is individual belief changes. Analyzing individual belief changes of participants of the great Devonian controversy, the following shows:

- ($H3.1$) Participants do not only change their dating hypotheses, but also their evidential beliefs. Often, participants hold on to a certain dating hypothesis while changing evidential beliefs. Given that participants are rational, this result dis-confirms strict falsificationism in the sense of Popper (1935).
- ($H3.2$) For the great Devonian controversy, there are dating hypotheses as well as evidential beliefs, which are constantly kept, or at least only very reluctantly given up. Hence, these dating hypotheses and evidential beliefs can be considered as hard core assumptions in the sense of Lakatos (1970).
- ($H3.3$) Most of the time, dating hypotheses as well as bodies of evidence are only slightly altered. This illustrates Laudan (1984) stating that beliefs are not revised as a whole, but rather in a piecemeal and reluctant way.

Some words on $H4$, that is rational belief change. Is rationality in belief change related with some kind of evidential support? Here, evidential support is spelled out in terms of $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. The following principles of rational belief change are tested:

- ($RAT1$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it does not decrease the dating hypothesis's degree of evidential support.
- ($RAT2$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it maximizes the dating hypothesis's degree of evidential support.

Here, a person maximizes her dating hypothesis's evidential support, if she chooses a dating hypothesis such that, given her body of evidence, there is no better confirmed dating hypothesis. For the great Devonian controversy, the following shows:

- ($H4.1$) Most of the time, individual belief changes are rational. However, there are individual belief changes which are not rational. Using $F_{DOJ}(h,e)$, individual belief changes are *more* often rational than using one of the other two confirmation measures.

- ($H4.2$) Shared belief changes are *less* often rational than individual belief changes. Using $F_{DOJ}(h, e)$, shared belief changes are *more* often rational than using one of the other two confirmation measures.

Presuppose that participants of the great Devonian controversy are rational. Together with $H4.1$, it follows that there are exceptions to the two previously introduced principles of rational belief change. This confirms Feyerabend (1976) stating that there is no scientific rule without any exceptions. However, it does not support relativism in the sense of "anything goes".

Some words on $H5$, that is consensus formation. Roads to the final consensus differ remarkably. There are no two persons following the same road. However, there are rather strong similarities. For all main participants of the great Devonian controversy, it holds:

- ($H5.1$) Approaching alternates with distancing the final consensus. This is in line with Betz (2013) investigating consensus-conduciveness of controversial debates by means of multi-agent simulations.
- ($H5.2$) Approaching the final consensus in terms of dating hypotheses does not imply an approachment in terms of bodies of evidence, and vice versa. This is a refinement of the analysis of consensus formation given in (Rudwick, 1988).

As a result of my analyses, it does not hold that a belief change distances the final consensus, iff it decreases evidential support. This is true for all three confirmation measures. However, it shows that, after a sufficiently large number of successive belief changes decreasing evidential support, there is often a considerable change in similarity with the final consensus. Hence, successive decrease of evidential support seems to be a reason for changing beliefs. Note that there are reasons for not changing beliefs, even if evidential support decreases.

As a further result of my analyses, it does not hold that a belief change distances the final consensus iff it does not maximize evidential support. This is true for all three confirmation measures. Further, it does not hold that a belief change approaches the final consensus, if it maximizes evidential support.

Is maximizing evidential support well designed to approach the final consensus, that is, do most dating hypotheses maximizing evidential support have a sufficiently high degree of similarity with the final dating hypothesis? For every time step and

person, the ratio of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize evidential is determined. There are several things about this ratio which should be noted:

- ($H5.3$) There are big differences between persons. Hence, there is a dependence between this ratio and a person's body of evidence. However, it cannot be assessed in terms of a body's similarity with the final consensus. A certain degree of similarity is neither a sufficient nor a necessary condition for the ratio being greater 0.5. Hence, some evidential claims seem to have more impact than others.
- ($H5.4$) There are differences between confirmation measures. In order to maximize the ratio, $Z_{DOJ}(h, e)$ is better than the other two confirmation measures.
- ($H5.5$) For all three confirmation measures and persons, before time step $4a$, it holds: The ratio is less or equal 0.5. Hence, during early phases of the debate, maximizing evidential support is not well designed to approach the final consensus, independent of a certain person.
- ($H5.6$) For all three confirmation measures and most main participants, from time step $7b$ till the end, the ratio equals 1. At the final step, this is true for all persons. Hence, the debate is closed when, as a result of argumentation, evidence accumulation and belief changes, maximizing evidential support is maximally well designed to approach the final dating hypothesis, independent of a certain person.

The two latter results newly illustrate a model of the closure of major scientific debates embodying ideas of rationalism as well as anti-rationalism, namely the compromise model as introduced in (Kitcher, 1993). Presume that (i) participants of the great Devonian controversy undergo the process of maximizing evidential support and (ii) cognitive progress is considered as approaching the final dating hypothesis. Then, $H5.5$ confirms condition $C4$ of the compromise model stating that, "[d]uring early phases of scientific debate, the processes undergone by the ultimate victors are (usually) no more well designed for promoting cognitive progress than those undergone by the ultimate losers" (Kitcher, 1993, p. 201). Further, presuming the same two assumptions, $H5.6$ confirms condition $C5$ of the compromise model, roughly stating that scientific debates end, as a result of argumentation, evidence accumulation and belief changes, when a certain cognitive process, which is executable for all participants, performs better in terms of cognitive progress than all the others

undergone by participants of the debate.

This section is organized as follows: The great Devonian controversy is shortly introduced and characteristics of its reconstruction are outlined in sec. 3.1, referring to common concepts in philosophy of science as theory-ladenness of observational data, Duhemean underdetermination, mini- and maxi-theories as well as newly introduced concepts as atomic dating hypotheses, shared background beliefs, bodies of evidence and two kinds of time slicing. Next, in sec. 3.2, polarization and polarization dynamics are assessed in terms of groups which are defined endogenously or exogenously. Further, individual belief changes are contrasted with some prominent philosophical views on this topic. For every main participant, sec. 3.3 presents confirmation dynamics. Results are compared for three different notions of confirmation and belief changes analyzed in terms of confirmation. Finally, in sec. 3.4, for every main participant, roads to the final consensus are analyzed. Special focus lies on relations between similarity with the final consensus and confirmation as well as the reliability of a certain kind of forming beliefs according to confirmation and its reliability defeaters.

3.1. Reconstruction a Historic Scientific Debate

Relying on (Rudwick, 1988), the great Devonian controversy is reconstructed as a series of dialectical structures. A dialectical structure consists of theses and deductive valid arguments possibly attacking and supporting one another. For more information on dialectical structures see (Betz, 2010) or sec. 2.1.2. There are several benefits of this reconstruction method. First, it reveals auxiliary assumptions and hypotheses, inferential relations and argument types. Second, it connects to computational analyses since a dialectical structure is a boolean formula.

Here, Argdown, as developed by Christian Voigt (2018), is used to implement and output dialectal structures. On these outputs, computational analyses are performed, using the computer algebra system Mathematica from Wolfram Research, Inc. (2019) and computing resources from Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT) (2017) and KIT's Department of Philosophy, ITZ and ITAS (2017).

Some words on observations and theories. Reconstructing the great Devonian controversy, the following shows:

- ($H1.1$) For all empirical statements of the debate, it holds: There is a dependence between its theoretical context and rational acceptance. Hence, the debate illustrates nicely the concept of theory-ladenness, which is uncontroversial in today's philosophy of science, see for example (Boyd and Bogen, 2021).
- ($H1.2$) An empirical statement is implied by some mineralogical or fossil criterion, only if it is conjoined with some auxiliary assumptions. Therefore, the debate illustrates Duhemian underdetermination, compare (Duhem, 1954).
- ($H1.3$) There are several criteria and most of them are highly controversial most of the time. Only at the end, there is a criterion which all participants agree upon. Therefore, the great Devonian controversy illustrates the struggle about standardizing methodological rules for generating empirical statements.
- There are only a few sentences which are accepted straight away by all participants all of the time, that is without being inferred. Here, these sentences are called *shared background beliefs*. Additionally, for every time step and participant, there are some more sentences which are accepted straight away. Together with the shared background beliefs, they form the so-called *body of evidence*. During the great Devonian controversy, for every participant, the body of evidence changes.

My reconstruction of the great Devonian controversy does not include principles of evidential support. What about goals? Certainly, every participant pursues his own goals. Quite certainly, all participants pursue at least one common goal, namely explaining as much phenomena as possible while only relying on generally accepted principles. However, goals are not part of my reconstruction either.

Some words on reconstructing in a trouble zone. Calculating degrees of confirmation, computing time exponentially increases with increasing number of sentences. For the great Devonian controversy, it shows that computing time is no longer reasonable, if a dialectical structure consists of more than 132 sentences. Therefore, reconstructing the great Devonian controversy, simplifications, omissions and work-around solutions are necessary.

The most important work-around solution is introducing atomic dating hypotheses. All the older strata of Devon are partitioned into three parts, namely the main part of the Culm strata, the black Culm limestone and the Non-Culm strata. For every part, the set of possible geological ages is the same: Cambrian, Silurian, Old Red

Sandstone, Mountain Limestone and Coal Measures. As a consequence, there are 15 so called *atomic dating hypotheses.* A *dating hypothesis* is a complete position on the set of atomic dating hypotheses. All mathematically possible combinations of these 15 atomic dating hypotheses are called *space of dating hypotheses.*

Using atomic dating hypotheses is advantageous for several reasons. First, every dating hypothesis can figure in an argument without introducing any new sentence. Second, degrees of confirmation are much more reliable using atomic dating hypotheses. The more dating hypotheses are taken into account, the more reliable are degrees of confirmation. Using atomic dating hypotheses, the whole space of dating hypotheses is taken into account. However, in order to be historically adequate, this space has to be restricted. Restriction is done via arguments connecting atomic dating hypotheses with one another as well as other sentences. For example, there are arguments forbidding dating hypotheses which contradict shared background beliefs.

Some words on time slicing. First, the time span of the great Devonian controversy is discretized, that is, the narrative as unfolded in (Rudwick, 1988) is divided into 9 successive episodes. Every episode is reconstructed as a dialectical structure. Moving forward in time, statements as well as arguments accumulate. The dialectical structure at time step $(t+1)$ augments the dialectical structure at time step t with some new sentences, arguments and inferential relations. Second, in order to analyze individual belief dynamics, a second slicing is performed separating changes in the *body of shared background beliefs* from others. The second time slicing splits each time step into two introducing t_a and t_b. Moving from time step $(t-1)_b$ to t_a, there is a new dialectical structure and there are some new *shared background beliefs.* However, individual beliefs stay the same. Moving from time step t_a to t_b, it is the other way around. The dialectical structure and body of *shared background beliefs* stay the same. However, there are some changes in individual beliefs.

3.1.1. Observations and Theories

The great Devonian controversy is a scientific debate. Among philosophers of science, opinions differ on what scientific debates are all about. Are they about accepting some general statements, mini-theories or maxi-theories? What are these things and is it possible to compare them among each other?

As for example Laudan (1977) notes, the term "theory" is used, at least, in two ways. First, it denotes a set of statements or models which are used to predict and explain phenomena. Examples of such mini-theories are Einstein's theory of the photoelectric effect and Maxwell's theory of electromagnetism. Second, the term "theory" denotes something much more general and less easily testable than a single mini-theory. Examples of such maxi-theories are the quantum theory or the kinetic theory of gases. Note also, that, as Laudan (1977, p. 72) puts it, "not only are there contrasts of generality and specificity between [these two types of theories], but the modes of appraisal and evaluation appropriate to each are radically different."

Some philosophers think of mini-theories as sets of statements, which are deductively organized and inferentially connected. Compare for example the so-called Zweistufenkonzept in (Schurz, 2006). According to the Zweistufenkonzept, a mini-theory comprises three different kinds of statements: Observational statements and two kinds of theoretical statements, namely empirical generalizations and principles. Others think of mini-theories from a non-statement point of view. There are such different non-statement point of views as those of Wolfgang Stegmüller (1969) and Giere (2010), thinking of mini-theories as set-theoretical models and model families, respectively.

Kuhn (1983) promote a very prominent view on maxi-theories. There, a maxi-theory is called a paradigm and comprises such different things as single mini-theories, epistemic values, ontological assumptions as well as paradigmatic examples of problems and their solutions. Later Laudan (1984) and Kitcher (1993) extended this account, both in their own way, speaking of "world views" and "practices", respectively. A "world view" consists not only of mini-theories, but also of ontological and methodological claims as well as aims. According to Laudan (1984), methodology comprises not only algorithms generating facts, but also principles of empirical support and comparative theory assessment. According to Kitcher (1993), every scientist possesses a so-called "individual practice" consisting of a certain language, a set of significant questions, explanatory and inferential patterns, criteria of credibility for information as well as reliability for experimentation. A "consensus practice" of a scientific community not only comprises the intersection of all individual practices, but the acknowledgment of authorities and an organization into sub-communities. A scientific debate is a competition between individual practices, ending after sufficient modifications of the corresponding consensus practice. All participants of a scientific debate pursue a common goal, namely the "production of a maximally uni-

fied set of explanations for the broadest possible class of phenomena" (see Kitcher, 2000, p. 29).

What is my reconstruction of the great Devonian controversy all about? Being a dialectical structure, my reconstruction is a set of sentences, deductively organized and inferentially interconnected. Therefore, my reconstruction is not about mini-theories from a non-statement point of view.

What kind of statements are there? Empirical and theoretical statements? What are empirical statements? As a starting point, empirical statements are those only using observational terms. However, at least for the great Devonian controversy, the meaning of an observational term is never fully determined by perceptual experiences. Therefore, for all empirical statements of this debate, it holds: There is a dependence between its theoretical context and rational acceptance. Hence, the debate illustrates nicely the concept of theory-ladenness, which is uncontroversial in today's philosophy of science, see for example (Boyd and Bogen, 2021). This dependence makes classification variable. As examples, consider the following sentences:

- ($E1$) Near Bideford, there are Coal Measures fossil plants in the main part of the Culm strata.
- ($E2$) Non-Culm strata are of characteristic Cambrian rock type.
- ($E3$) The passage between the main part of the Culm strata and the black Culm limestone is conformable.
- ($E4$) In strata older than Old Red Sandstone strata, there are no Carboniferous fossils.
- ($E5$) The black Culm limestone is older than the main part of the Culm strata.

Clearly, $E1$ states an observation, namely some finding at a certain location. However, it takes a lot of theoretical context to classify this finding as fossil plants being Coal Measures in age. $E2$ also seems to state an observation, namely a certain appearance of strata. However, not only the classification of this appearance as characteristic Cambrian relies on theoretical assumptions, as for example the ontological assumption that there is a characteristic Cambrian rock type. Moreover, $E2$ is a generalization from a limited set of observations, as well as $E3$.[1] Perhaps, $E3$ is even only an approximation neglecting some locations where the passage is unconformable. $E4$ seems to be an empirical statement stating an observation, namely

[1]Here, I take it that it is impossible to really observe *all* strata in Devon.

that no Carboniferous fossil has been found in any stratum older than Old Red Sandstone. However, Murchison does not accept $E4$ after having studied all strata older than Old Red Sandstone. Rather, he infers $E4$ from his other beliefs and some minor fieldwork in a small corner of the Welsh borderland. $E5$ does not seem to use an observational term at all. However, given a mineralogical or fossil criterion and some auxiliary assumptions, age can be an observational term as much as for example time, current, voltage, force, pressure, entropy or temperature.

During the great Devonian controversy, participants infer different hypotheses about the age of all the older strata in Devonshire from evidential beliefs which often change. These inferences are based on so-called mineralogical and fossil criteria connecting the mineralogical character and fossil content of certain strata with certain geological ages, respectively. For the great Devonian controversy, there are several criteria. Consider the following examples:

- ($E6$) Strata are originally formed from sediments that were deposited in flat horizontal sheets - the younger sediments deposited on older ones.
- ($E7$) For two strata A and B it holds: (i) The more A and B are similar in their fossil assemblages, the more they are similar in age, (ii) the more A and B are similar in age, the more they are similar in their fossil assemblages and (iii) if A and B are sufficiently unsimilar in age, then they have no species in common.
- ($E8$) There is a bijection between the age of some strata and their characteristic fossil assemblages.
- ($E9$) There is a bijection between the age of some strata and their characteristic rock types.
- ($E10$) Given a sufficiently large (i) amount of fossils and (ii) region under study, it holds: There is a bijection between the age of some strata and their characteristic fossil species assemblages.

Given that strata are undisturbed, $E6$ serves as a criterion translating relative position into relative age. $E8$ is implied by $E7$ which is known as Lyell's principle stating a constant piecemeal change in fauna and flora. $S8$ and $S9$ translate fossil assemblages and rock types into age, respectively. Let us take a closer look at $E9$. Together with assumptions about their age and its corresponding characteristic rock type, $E9$ implies the rock type of some strata at a certain location. Together with

assumptions about their rock type and characteristic rock types, $E9$ implies the age of some strata at a certain location. $E10$ is a new version of $E8$ limiting its scope.

All those criteria are general statements inferentially connecting dating hypotheses with empirical statements and auxiliary assumptions. Hence, for the great Devonian controversy, it shows that an empirical statement is implied by some mineralogical or fossil criterion, only if it is conjoined with some auxiliary assumptions. Therefore, the debate illustrates Duhemian underdetermination, compare (Duhem, 1954). Due to mineralogical and fossil criteria being bijections, a dating hypothesis is implied by some empirical statement and some auxiliary assumptions.

Further, the great Devonian controversy illustrates the struggle about standardizing methodological rules for generating empirical statements. Most of the criteria are highly controversial most of the time. Only $E6$ is uncontroversial all of the time. However, an empirical statement is implied by $E6$, only if strata are undisturbed. This auxiliary assumption is often controversial. Only at the end, there is a criterion, which all participants agree upon, namely $E8$. Most mineralogical and fossil criteria are attacked by arguments based on the assumption that there are always local variations. Due to the meaning of 'characteristic', there are characteristic fossils and fossil assemblages, only if there are no local variations in flora and fauna. For the same reason, there are characteristic rock types, only if there are no local variations in sedimentation. Therefore, at least for this certain debate, classifying all those statements as empirical,which are generated by some standardized methodological rule does not seem to be pragmatic.[2]

So, what is a pragmatic way of dealing with this variability in classification? Classifying all those sentences as empirical, which are accepted straight away, that is without being inferred? Classifying all those sentences as empirical which are accepted by all participants all of the time? Or a combination of both? During the great Devonian controversy, there are only a few sentences which are accepted straight away by all participants all of the time, that is without being inferred. Here, these sentences are called *shared background beliefs*. See Fig. B.4 for a detailed listing of all shared background beliefs. Additionally, for every time step and participant, there are some more sentences which are accepted straight away. Together with the shared background beliefs, they form the so-called *body of evidence*. See sec. B.3 for a detailed listing of all bodies of evidence. From a *body of evidence* and a dialectical

[2]Compare (Chalmers, 2007) for some thoughts on observations, standardized observation methods and objectivity of observational statements.

structure, further beliefs are inferred. During the debate, for every participant, the body of evidence changes. However, these changes are considerably different, see for example sec. 3.2.2 as well as sec. 3.4.

How does a body of evidence support a dating hypothesis? My reconstruction of the debate does not include principles of evidential support. However, based on my reconstruction, three different Bayesian measures of evidential support are calculated and compared, see sec. 2.1.4. In terms of evidential support, principles of comparative dating hypotheses assessment are proposed. For the great Devonian controversy, these principles are used to explain belief dynamics of the main participants, see sec. 3.4.

What about cognitive aims and values? Certainly, every participant of the great Devonian controversy pursues his own goals. Consider the following examples:

- Murchison tries to promote the Silurian system, as defined by himself after some fieldwork in the Welsh borderland as a global system, striving for fame.
- Lyell is keen on finding evidential support for his own principle.
- De la Beche tries to vindicate his competence in the field in order to secure his livelihood. Further, De la Beche wants to promote a certain kind of science, namely a science free from preconceived opinions.

Individual aims are not part of my reconstruction. What about common goals? Here, I take it that all participants of the great Devonian controversy pursue at least one common goal, namely explaining as many phenomena as possible while only relying on generally accepted principles. Are there other common goals in terms of evidential support? For example, do all participants strive to maximize, or at least increase, evidential support of their dating hypotheses? For the great Devonian controversy, an answer to this question can be found in sec. 3.3.2.

3.1.2. Reconstructing in a Trouble Zone

Calculating degrees of confirmation, computing time increases exponentially with increasing number of sentences. For the great Devonian controversy, it shows that computing time is no longer reasonable, if a dialectical structure consists of considerably more than 132 sentences. Therefore, reconstructing this debate, simplifications and omissions are necessary. Nevertheless, in order to ensure informative results,

dissent and consensus have to be captured in an intelligible way. Hence, reconstructing the great Devonian controversy, one operates in a trouble zone. In the following, I will shortly summarize omissions, simplifications and work-around solutions.

Various things have been omitted. First, there are several sub-debates having been omitted, as for example the debates about the dating of certain strata in the Rhineland, the Eifel and North-west France. These debates have been very lively and controversial, only coming to an end by relying on the final consensus of the great Devonian controversy. Here, I take it that, omitting these sub-debates, dissent and consensus are nevertheless captured in an intelligible way. Second, there are several statements having been omitted, as for example fossil findings from Ireland, the French Alps and North America, namely fossils resembling those of the Coal Measures era but found in strata older than Old Red Sandstone. For more examples see sec. B.1. Finally, analyses are performed only for the six main participants, these are De la Beche (DLB), Murchison (MUR), Lyell (LYE), Phillips (PHI), Sedgwick (SED) and Austen (AUS). Omitted persons have beliefs which are at least very similar to those of a non-omitted person.

There are two kinds of simplifications, namely sentence merging and presupposing uncontroversial sentences. Clearly, sentence merging is advantageous, namely in diminishing the number of sentences. However, it can be disadvantageous as well, namely in diminishing the historical adequateness of a person's position. As an example, consider the following sentence: The fossil assemblage of the black Culm limestone is no local variation, that is, (i) the region under study as well as (ii) the amount of fossil species is sufficiently large. For a certain time interval, it seems historically quite adequate to suppose that De la Beche accepts part one and rejects part two. For more examples see sec. B.1. If uncontroversial throughout the debate, a sentence is presupposed that is, before any calculation, the corresponding truth value is set as true. In this thesis, such sentences are called *shared background belief*s.

Finally, there is a work-around solution, namely introducing atomic dating hypotheses. All the older strata of Devon are partitioned into three parts, namely the main part of the Culm strata (MC), the black Culm limestone (BCL) and the Non-Culm strata (NC). For every part, the set of possible geological ages is the same consisting of Cambrian, Silurian, Old Red Sandstone, Mountain Limestone and Coal Measures. As a consequence, there are 15 so-called *atomic dating hypotheses*, compare definition 3.1.1. Here, a *dating hypothesis* is a complete position on the set of atomic

dating hypotheses.[3] All mathematically possible combinations of these 15 atomic dating hypotheses constitute the *space of dating hypotheses*.

Definition 3.1.1: Atomic Dating Hypotheses

An atomic dating hypothesis is a sentence of the following type:
Some part of strata x is y in age.
Here, x is some part of the older strata of Devon, that is the main part of the Culm, the black Culm limestone or Non-Culm strata, and y is one of five possible geological ages, namely Cambrian, Silurian, Old Red Sandstone, Mountain Limestone or Coal Measures.

Using atomic dating hypotheses is advantageous for several reasons. First, every dating hypothesis can figure in an argument without introducing any new sentence. Second, degrees of confirmation are much more reliable using atomic dating hypotheses. The more dating hypotheses are taken into account, the more reliable are degrees of confirmation. Using atomic dating hypotheses, the whole space of dating hypotheses is taken into account. However, in order to be historically adequate, this space has to be restricted. Restriction is done via arguments connecting atomic dating hypotheses. Some of these arguments introduce new sentences. That is why some of them have to be omitted. So, the space of atomic dating hypotheses is only approximately historically adequate.

First, there are arguments forbidding dating hypotheses which contradict one of the following shared background beliefs:

- ($E11$) The main part of the Culm strata is conformable, that is, there is no gap in its temporal sequence.
- ($E12$) The black Culm limestone is conformable, that is, there is no gap in its temporal sequence.
- ($E13$) The black Culm limestone is older than the main part of the Culm strata.
- ($E14$) The passage between the main part of the Culm strata and the black Culm limestone is conformable.
- ($E15$) Culm strata are not intercalated by Non-Culm strata.

[3] For some explanatory notes on the relation between *dating hypotheses* and *interpretative schemes* as defined in (Rudwick, 1988, p. 407) see sec. B.2.

In the following, an example of an argument connecting some atomic dating hypotheses with sentence $E11$ is presented. Given that there is no gap in its temporal sequence, no part of the main part of the Culm is Coal Measures in age, if some part is Old Red Sandstone and no part is Mountain Limestone in age.

Example 3.1.1: Argument type: Conformable formation

(P1) Given strata are undisturbed, Primary strata are overlain by Cambrian strata, Silurian strata, Old Red Sandstone strata, Mountain Limestone strata, Coal Measures strata, New Red Sandstone strata and Oolitic stata, respectively.
(P3) Some part of the main part of the Culm strata is Old Red Sandstone in age.
(P4) No part of the main part of the Culm strata is Mountain Limestone in age.
(E11) The main part of the Culm strata is conformable, that is, there is no gap in its temporal sequence.
—
(C) No part of the main part of the Culm strata is Coal Measures in age.

Note that $P1$ is a shared background belief defining the standard sequence of geological ages. For more shared background beliefs see Fig. B.4. Consider also the following argument connecting some atomic dating hypotheses with sentence $E15$. It states that, if the main part of the Culm strata encompasses Coal Measures as well as Old Red Sandstone strata, then there are no Non-Culm strata which are Mountain Limestone in age.

Example 3.1.2: Argument type: Culm not intercalated

(P1) Given strata are undisturbed, Primary strata are overlain by Cambrian strata, Silurian strata, Old Red Sandstone strata, Mountain Limestone strata, Coal Measures strata, New Red Sandstone strata and Oolitic stata, respectively.
(P2) Some part of the main part of the Culm is Coal Measures in age.
(P3) Some part of the main part of the Culm is Old Red Sandstone in age.
(E15) Culm strata are not intercalated by Non-Culm strata.
—
(C) No part of the Non-Culm strata is Mountain Limestone in age.

Second, restriction is done via arguments connecting atomic dating hypotheses with sentences of the following type:

- ($E16$) The passage between strata V and strata W is conformable.
- ($E17$) In Devon, strata are in the temporal order X.
- ($E18$) In strata of age Y, there are no fossils of type Z.

These sentences are controversial throughout the great Devonian controversy. As already shortly explained in the previous section, their classification as empirical is contextual. An argument connecting some anatomic dating hypotheses with a sentence of type $E16$ is presented in the following. Given a conformable passage between the black Culm limestone and Non-Culm strata and certain atomic dating hypotheses, it follows a certain atomic dating hypothesis. If the Black Culm limestone, passing conformably into Non-Culm strata, is at oldest Mountain Limestone and there are no Non-Culm strata which are Mountain Limestone in age, then there are some Non-Culm strata which are Old Red Sandstone in age.

Example 3.1.3: Argument type: Conformable passages

(P1) Given strata are undisturbed, Primary strata are overlain by Cambrian strata, Silurian strata, Old Red Sandstone strata, Mountain Limestone strata, Coal Measures strata, New Red Sandstone strata and Oolitic stata, respectively.
(P2) Some part of the black Culm limestone is Mountain limestone in age.
(P3) No part of the black Culm limestone is Old Red Sandstone in age.
(P4) No part of the Non-Culm strata is Mountain Limestone in age.
(E16*) The passage between the black Culm limestone and the Non-Culm strara is conformable.
—
(C) Some part of the Non-Culm strata is Old Red Sandstone in age.

Consider also the following argument connecting some atomic dating hypotheses with a sentence of type $E17$. It states that, if the main part of the Culm strata is older than some Non-Culm strata, being at youngest Old Red Sandstone in age, then the main part of the Culm strata is at youngest Old Red Sandstone.

Example 3.1.4: Argument type: Older than

(P1) Given strata are undisturbed, Primary strata are overlain by Cambrian strata, Silurian strata, Old Red Sandstone strata, Mountain Limestone strata, Coal Measures strata, New Red Sandstone strata and Oolitic stata, respectively.
(P2) Culm strata and Non-Culm strata are both older than Primary and younger than New Red Sandstone.
(P3) No part of the Non-Culm strata is Coal Measures in age, and no part of the Non-Culm strata is Mountain Limestone in age.
(E17*) Some Non-Culm strata are younger than the main part of the Culm.
—
(C) No part of the main part of the Culm strata is Coal Measures in age, and no part of the main part of the Culm strata is Mountain Limestone in age

Note that $P2$ is a shared background belief stating that there are 5 possible geological ages for all the strata which should be dated. For more shared background beliefs see Fig. B.4.

Finally, there are arguments connecting atomic dating hypotheses with mineralogical and fossil criteria. Given a criterion, a dating hypothesis and possibly some auxiliary assumptions, an observation can be made. As already shortly indicated in the previous section, the corresponding empirical statement hinges on ontological assumptions. Possible auxiliary assumptions are definitions of characteristic rock types, fossils and fossil assemblages. In case of a limited scope, there are auxiliary assumptions about the applicability of the criterion. As an example, consider the following argument. If strata can be identified by means of their characteristic fossils and the main part of the Culm strata is Coal Measures in age, then it supports characteristic Coal Measures fossils.

Example 3.1.5: Argument type: Criterion

(P1) There is a bijection between the age of some strata and its characteristic fossils.
(P2) The main part of the Culm strata is Coal Measures in age.
—
(C) There are characteristic Coal Measures fossils in the main part of the Culm strata.

In my reconstruction of the great Devonian controversy, arguments and sentences are clustered according to their main theme or purpose. There are 10 such thematic clusters:

- *Observations*: This thematic cluster only comprises statements, that is there are no arguments at all. These statements can be considered as observational statements.
- *Dating of the main part of the Culm strata*: This cluster comprises atomic dating hypotheses with respect to the main part of the Culm strata. Here, there are arguments of type *"Criterion"* supporting certain atomic dating hypotheses with respect to the main part of the Culm strata, compare example 3.1.5. Further, there are arguments of type *"Conformable Formation"* spelling out that the main part of the Culm strata is a conformable formation, compare example 3.1.1.
- *Dating of the black Culm limestone*: This cluster comprises atomic dating hypotheses with respect to the black Culm limestone. Here, there are arguments of type *"Criterion"* supporting certain atomic dating hypotheses with respect to the black Culm limestone, compare example 3.1.5. Further, there are arguments of type *"Conformable Formation"* spelling out that the black Culm limestone is a conformable formation, compare example 3.1.1.
- *Dating of the Non-Culm strata*: This cluster comprises atomic dating hypotheses with respect to the Non-Culm strata. Here, there are arguments of type *"Criterion"* supporting certain atomic dating hypotheses with respect to the Non-Culm strata, compare example 3.1.5.
- *Criteria*: This cluster comprises mineralogical and fossil criteria as well as arguments attacking them. Attacks are supported for example by the existence of local variations in sedimentation or the existence of local variations in fauna and flora.
- *Youngest Devonian Strata*: What follows from a certain temporal ordering and certain atomic dating hypotheses? In this thematic cluster, there are several arguments answering this question, for several different temporal orderings as well as atomic dating hypotheses.
- *Gap in the sequence*: What follows from a certain temporal ordering and certain atomic dating hypotheses, if one additionally presupposes certain geological conformities? In this thematic cluster, there are several arguments

of type *"Conformable Passages"* answering this question, for several different temporal orderings, atomic dating hypotheses as well as geological conformities, compare example 3.1.3. Here, the term "geological conformity" is used to denote two strata passing conformably into one another, that is the non-existence of a gap in the geological sequence.

- *Carboniferous fossils in strata older than Old Red Sandstone:* This cluster is all about the existence of Carboniferous plants in strata older than Old Red Sandstone. From the beginning, Lyell's principle attacks this claim. Together with fossil findings in some north as well as south Devonian Non-Culm strata, denying this claim rules out several atomic dating hypotheses with respect to the Non-Culm strata.
- *Old Red Sandstone characteristics:* What rock types, fossils or fossil assemblages are characteristic of Old Red Sandstone strata? Throughout the Great Devonian controversy, this question is highly controversial. In the beginning, Scottish Old Red Sandstone strata serve as a blueprint, both in rock type and fossils. However, Lyell's principle supports a hypothesis which cannot be reconciled with the fossil content of Scottish Old Red Sandstone strata.
- *Other regions than Devon:* This cluster comprises arguments using observations from other regions than Devon, as for example Scotland, Yorkshire, Pembrokeshire and Russia.

For every thematic cluster, Fig. 3.1 shows the number of arguments over time.

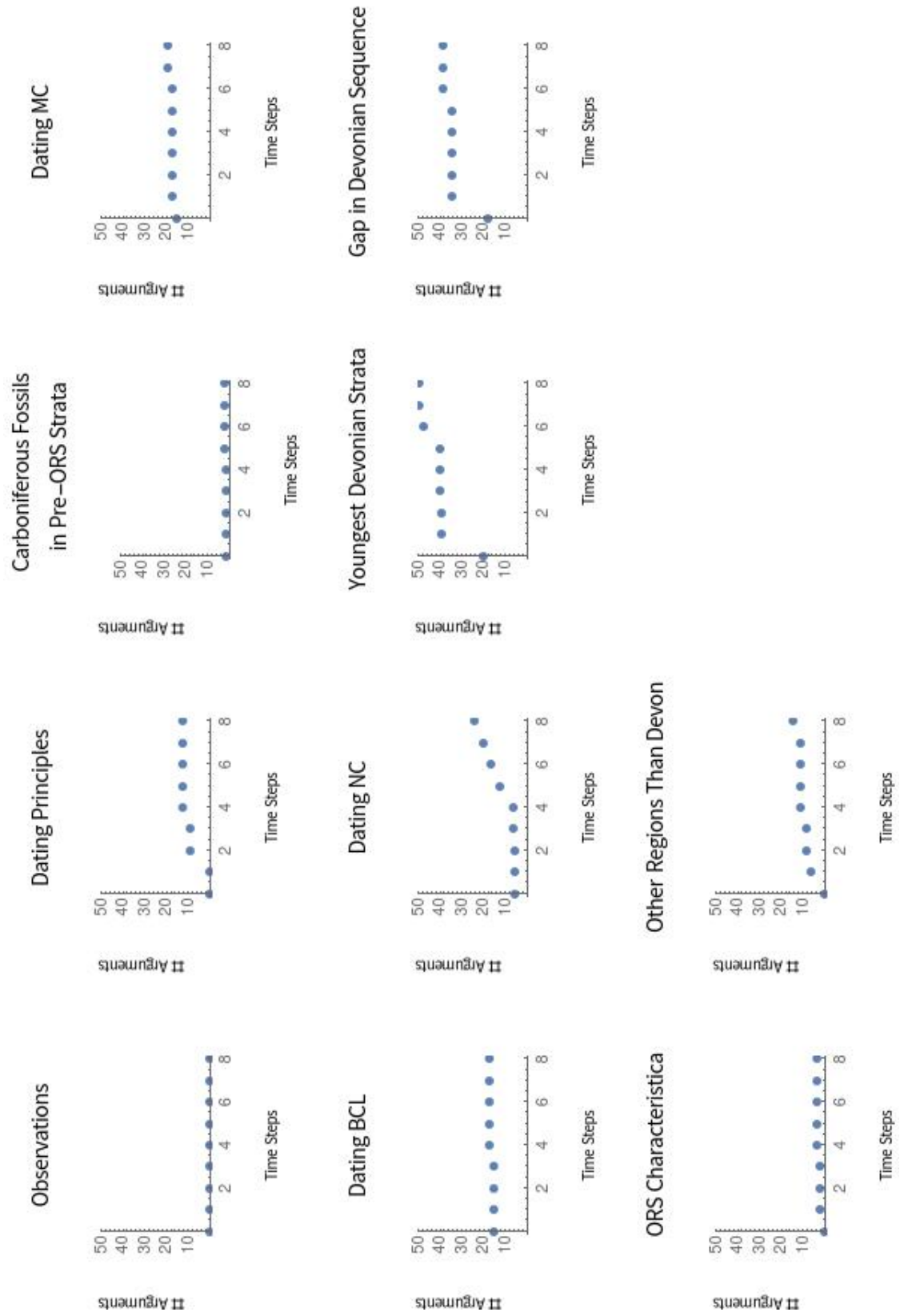

Figure 3.1.: Argument Dynamics. Arguments of the great Devonian controversy are clustered according to their main theme. For every cluster, there is a plot showing the number of arguments over time.

3.1.3. Two Kinds of Time Slicing

The aim of this thesis is to analyze scientific reasoning with fallible evidence in terms of confirmation. I am especially interested in relations between belief dynamics and confirmation dynamics. Therefore, two time slicings are indicated.

First, in order to observe dynamics, the time span of the great Devonian controversy is discretized, that is, the narrative is divided into successive episodes. Here, time is discretized into 9 time steps, namely $0, 1, ..., 8$. Fig. 3.2 shows this discretization on top of a schematic chart of the development of the great Devonian controversy taken from (Rudwick, 1988). For every time step, the episode of the narrative is reconstructed as a dialectical structure. For some information on the theory of dialectical structures see sec. 2.1.2. Moving forward in time, statements as well as arguments accumulate. The dialectical structure at time step $(t + 1)$ augments the dialectical structure at time step t with some new sentences, arguments and inferential relations.

For every time step, the inferential density of the dialectical structure is calculated. See sec. 2.1.2 for its definition. Fig. 3.3 shows that the inferential density does not constantly increase with time. Rather, it is maximal and minimal at time step 4 and 1, respectively. Note that the inferential density is always larger than 0.74, which is remarkably large and much larger than inferential densities in the first part of this thesis.

Second, in order to analyze individual belief dynamics, a second time slicing is performed separating changes in the *body of shared background beliefs* from others. The second time slicing splits each time step into two, introducing t_a and t_b with $t \in \{0, ..., 8\}$. Moving from time step $(t-1)_b$ to t_a, there is a new dialectical structure and there are some new *shared background beliefs.* However, individual beliefs stay the same. Moving from time step t_a to t_b, it is the other way around. The dialectical structure and body of *shared background beliefs* stay the same. However, there are some changes in individual beliefs.

Based on Rudwick (1988) and my reconstruction, for every main participant and time step, a body of evidence (e) and a dating hypothesis (h) of all the older strata in Devon are identified, see sec. B.3 and Fig. B.3. Relying on the corresponding dialectical structure, further beliefs can be inferred from the body of evidence and the dating hypothesis. Altogether, they constitute a person's position at a certain time step.

In the following, I want to summarize very briefly the main points for every episode. Time step 0 is all about De la Beche. Based on own fieldwork as well as previous work of other acknowledged geologists, he delineates the geological structure of Devon, while being at a meeting of English geologists. Note that he considers all the older strata of Devon as Cambrian in age. Here, I take it that, at time step 0, De la Beche implicitly accepts dating strata by means of characteristic rock types. The next episode, time step 1, is about Murchison and Lyell immediately responding to De la Beche. They both reject the characteristic rock type criterion, but for different reasons. Murchison only accepts dating by means of characteristic fossils. Lyell, promoting his own principle, only accepts dating by means of characteristic fossil assemblages. Applying his criterion and assuming some of De la Beche's fossil findings to be characteristic Coal Measures fossils, Murchison claims the main part of the Culm strata to be Coal Measures in age. This way, he avoids giving up one of his central beliefs, namely that there are no Carboniferous fossils in strata older than Old Red Sandstone. This belief is also held by Lyell as a consequence of his principle. At time step 2, Phillips attacks all yet proposed mineralogical and fossil criteria by claiming that there are always variations in fauna and flora as well as in sedimentation. Thereby, he rejects the existence of characteristic fossils, fossil assemblages and rock types. As part of the same episode, De la Beche attacks Murchison's central belief by citing observations from other regions and countries. Here, I take it that De la Beche implicitly accepts all of Phillips's objections and, as a consequence, changes his beliefs. The next episode, time step 3, is all about the Devon campaign of Murchison and Sedgwick. At a meeting of English geologists in Bristol, they present their results, namely a new geological structure and some minor changes in Murchison's former dating hypothesis. Observing a Culm trough but no trough around Exmoor, Murchison and Sedgwick infer the Culm strata to be at the top - not at the middle part - of the sequence. Relying on some fossil findings, Murchison dates some Non-Culm strata as Silurian. The next episode, time step 4, is about Phillips immediately responding to Murchison and Sedgwick. His reasoning is in line with his former beliefs. First, he confirms variations in sedimentation with observations from the Pennines. Second, Phillips introduces a new fossil criterion, more precisely, a limited version of dating by means of fossil assemblages: Given a sufficiently large amount of fossils and a sufficiently large region under study, strata can be dated by means of fossil assemblages. Based on his new criterion as well as some observations around Yorkshire, Phillips infers the black Culm limestone to

be Mountain Limestone in age. The next episode, time step 5, is about new fossil findings in North Devon. There, Carboniferous fossil plants are found at various locations in Non-Culm strata. These new findings are problematic for all those considering the Non-Culm strata as older than Old Red Sandstone and, at the same time, denying the existence of Carboniferous fossils in such strata. The situation gets even worse at the next episode, time step 6, centering about new fossil findings in South Devon. There, some Non-Culm strata support shells and corals, which are at least very similar to those of the Carboniferous. Here, De la Beche goes one step back, placing the Culm strata once again in the middle of the sequence. For him, this move is rewarding in a twofold way. De la Beche can join the growing consensus on the Old Red Sandstone dating of South Devonian shells and corals. Together with Carboniferous fossils in the main part of the Culm strata, De la Beche disconfirms Murchison's central belief, namely that there are no Carboniferous fossils in strata older than Old Red Sandstone. At time step 7, in order to resolve his problems, Murchison revises his beliefs fundamentally introducing a new dating of the Non-Culm strata, namely assuming the whole Non-Culm strata to be Old Red Sandstone. Thereby, he assumes a fossil assemblage corresponding to the Old Red Sandstone to be intermediate between those of the Silurian and the Mountain Limestone. However, his argumentation relies on a comparatively small amount of fossil findings. Further, Murchison's argumentation makes fossils found in Scottish Old Red Sandstone strata a local variation which is a rather unusual assumption at that time. During the same episode, based on some additional fieldwork, De la Beche presents a new geological structure of Devon supporting his dating hypothesis. The final episode, time step 8, summarizes two big campaigns, both important for achieving the final consensus. Phillips' fieldwork on the Non-Culm strata neatly documents a constant piecemeal change in fauna and flora ranging from Silurian to Mountain Limestone fossils. Murchison campaigns Russia finding strata sandwiched by Silurian and Mountain Limestone strata and, at the same time, uniting fossils from Scottish Old Red Sandstone and the Non-Culm strata.

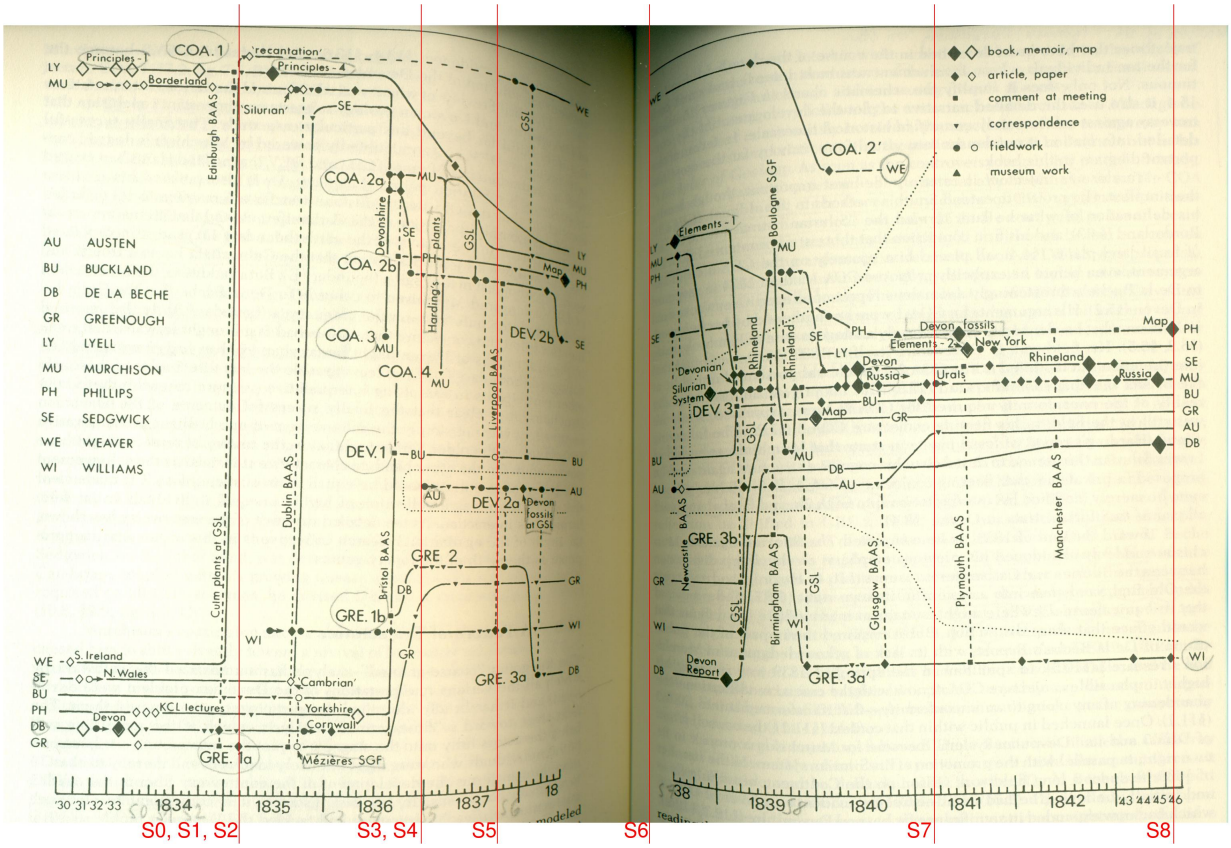

Figure 3.2.: Time slicing (red) overlying a schematic chart of the development of the great Devonian controversy taken from (Rudwick, 1988, p. 412). For ten participants, schemes for the interpretation of the older strata of Devonshire are sorted by time and so-called *theoretical distance*. Red lines indexed with several time steps are due to differences in the analyses, compare sec. B.2.

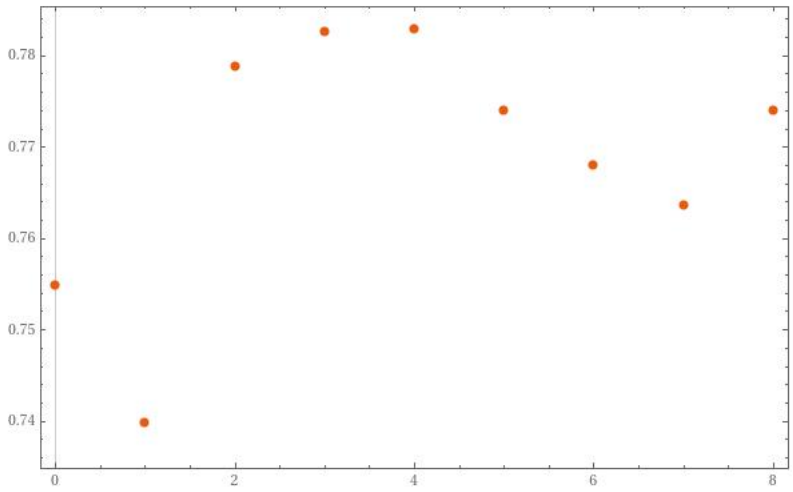

Figure 3.3.: Inferential Density Dynamics. This plot shows the inferential density of the great Devonian controversy for 9 successive time steps.

3.2. Polarization Dynamics

Rudwick (1988) recreates the great Devonian controversy as a narrative and delineates polarization and polarization dynamics as follows. First, there is a disagreement dividing participants into two parties facing each other rather irreconcilable. As the debate evolves, with new evidence popping up, some kind of middle ground seems to emerge. Finally, there is a consensus between most of the main participants which is no mere compromise but incorporates beliefs from both parties.

Here, polarization is assessed in terms of groups. They are not only defined exogenously but also endogenously, namely not only by acceptance of certain statements but also by mutual similarity between sets of statements. Bramson et al. (2017) propose a variety of polarization measures, as for example community fragmentation, distinctness, group consensus and size parity:

1. How many groups can be defined? (Community fragmentation)
2. How are participants distributed over groups? Are all groups more or less comparably sized? Or is there one dominant group? (Size parity)
3. To what extent do positions of members of the same group differ? (Group Consensus)
4. Are there shared beliefs between members of different groups? (Distinctness)

For practical reasons, the following analyses center on community fragmentation and size parity. However, using a similarity measure to define groups endogenously, group consensus as well as distinctness are assessed as well. The main findings of these analyses are:

- ($H2.1$) For exogenous clustering, groups are never the same as those in (Rudwick, 1988). This is not true for endogenous clustering. Clustering maximally similar dating hypotheses, groups are the same as those in (Rudwick, 1988).
- ($H2.2$) Similarity spectra of dating hypotheses and bodies of evidence are quite similar, namely $[0.60, 1.00]$ and $[0.54, 0.99]$. However, similarity dynamics of dating hypotheses and bodies of evidence do not coincide:
  - Except during the middle section, there are always some persons accepting the same dating hypothesis. Never, not even at the end, there are two persons accepting the same body of evidence.

  - Several times, there are two persons accepting, at the same time, remarkably similar bodies of evidence but unsimilar dating hypotheses, and vice versa.

- ($H2.3$) The average degree of similarity is maximal at the final step, not at last due to argumentation. In so far as dating hypotheses and bodies of evidence together constitute a paradigm, this result may be understood as an illustration of Kuhn (1983) stating that controversies are not only triggered but also resolved by inter-paradigmatic exchange of arguments.

Finally, this section concludes with some analyses of individual belief change. The following results should be noted:

- ($H3.1$) Participants do not only change their dating hypotheses, but also their evidential beliefs. Often, participants hold on to a certain dating hypothesis while changing evidential beliefs. Given that participants are rational, this result dis-confirms strict falsificationism in the sense of Popper (1935).

- ($H3.2$) For the great Devonian controversy, there are dating hypotheses as well as evidential beliefs which are constantly kept, or at least only very reluctantly given up. Hence, these atomic dating hypotheses and evidential beliefs can be considered as hard core assumptions in the sense of Lakatos (1970).

- ($H3.3$) Most of the time, dating hypotheses as well as bodies of evidence are only slightly altered. This illustrates Laudan (1984) stating that beliefs are not revised as a whole, but rather in a piecemeal and reluctant way.

3.2.1. Exogenously Defined Groups

Here, groups are defined exogenously, that is according to the acceptance of certain sentences, namely those which are called *interpretative boundaries* by Rudwick (1988). According to Rudwick (1988), in the beginning of the debate, there is a disagreement dividing participants of the great Devonian controversy into two parties facing each other rather irreconcilable. I*nterpretative boundaries* are "the banners flying over the battle lines". A listing of *interpretative boundaries* and their corresponding sentences in my reconstruction shows Fig. 3.4. For every main participant and *interpretative boundary*, the person's attitude towards the boundary is shown

in Fig. 3.5.[4]

	Sentence of the Reconstruction	Interpretative Boundary
B1	In strata older than Old Red Sandstone, there are no Carboniferous fossils	No Coal Measures plants in strata [older than Old Red Sandstone]
B2	In Devon, all the older strata form an unbroken sequence of strata	No gap in the sequence of Devonshire strata
B3	Some part of the Non–Culm strata is Old Red Sandstone in age.	[...] at least the younger part of the pre–Culm sequence [is] equated with the Old Red Sandstone
B4	Scottish Old Red Sandstone fauna and flora is a local variation.	the 'Devonian', equivalent with the Old Red Sandstone elsewhere, was regarded as a major system
B5	In Devon, among all the older strata, the main part of the Culm strata are not the youngest.	Culm [is placed] somewhere in the middle of the sequence rather than at the top

Figure 3.4.: Sentences of the reconstruction corresponding to *interpretative boundaries* as introduced in (Rudwick, 1988, p. 405).

Based on Fig. 3.5, groups are defined exogenously using definition 3.2.1. Results are shown in Fig. 3.6 and discussed in the following.

Definition 3.2.1: Clustering according to interpretative boundaries

Some persons form a group with respect to B_i, iff they exhibit the same attitude towards the interpretative boundary B_i.
Here, B_i with $i = 1, ..., 5$ is an interpretative boundary as listed in Fig. 3.4. Possible attitudes towards B_i are acception, rejection and judgment suspension.

First, some words on clustering according to *interpretative boundary* $B1$. From beginning to end, there are two groups, namely one group consisting of Murchison and Lyell and another group comprising all the others. Hence, $B1$ is not part of the final consensus.

[4]Some technical remarks: For some persons and time steps, there is an interpretative boundary,which is neither part of the dating hypothesis nor the body of evidence. Nevertheless, a person's attitude towards this boundary is determinable. First, based on the dialectical structure, dating hypothesis and body of evidence, further beliefs are inferred. Second, comparing these further beliefs with the interpretative boundary, the person's attitude towards the boundary is determined.

Second, some words on $B2$-groups. At first, there is only one group, namely Murchison and Lyell both rejecting $B2$. With Phillips entering the stage, there is a second one, namely Phillips and De la Beche both accepting $B2$. Entering the stage, Sedgwick first joins the group of Murchison and Lyell. At time step $4b$, Phillips changes groups. With Austen entering the stage, there is a new group of persons accepting $B2$, namely De la Beche, Austen and Sedgwick. At the next step, Murchison and Lyell join this group. Finally, Phillips also agrees on accepting $B2$.

Third, $B3$ is uncontroversial for quite a while. From the beginning until time step $5b$, all persons agree on rejecting $B3$. Even at time step $5b$, there is only one person, namely Murchison, accepting $B3$. Murchison changes his attitude towards $B3$ already at the next step. However, then, there are some other persons accepting $B3$, namely De la Beche and Austen. At time step $7b$, Murchison and Lyell join this group. At the end, all persons agree on accepting $B3$.

Fourth, $B4$ enters the stage rather late, namely at time step $4a$. Based on the dialectical structure and their bodies of evidence, there are two persons which have to reject $B4$, namely Murchison and Lyell. All other persons are free in their choice of attitude towards $B4$. De la Beche and Phillips choose to accept $B4$. Sedgwick joins Murchison and Lyell in rejecting $B4$. With Austen entering the stage, De la Beche and Phillips find another ally. At time step $7b$, Murchison revises his beliefs fundamentally including his attitude towards $B4$. Finally, all persons agree on accepting $B4$.

Fifth, among all persons but De la Beche, there is always an agreement on rejecting $B5$. De la Beche's attitude towards $B5$ changes constantly. First, he accepts $B5$. However, at time step $3b$, he joins the others in rejecting $B5$. Already at time step $6b$, De la Beche readopts his former belief. Finally, he reverses his attitude towards $B5$ once again.

Note that, for no clustering according to some *interpretative boundary*, groups are the same as those in Fig. 3.2 respectively (Rudwick, 1988). Reconstructing the great Devonian controversy as a dialectical structure, it shows that a person's position always comprises a lot more than five sentences. Therefore, clustering according to *interpretative boundaries* seems rather inadequate. The following subsection introduces another clustering algorithm designed to remedy this shortcoming.

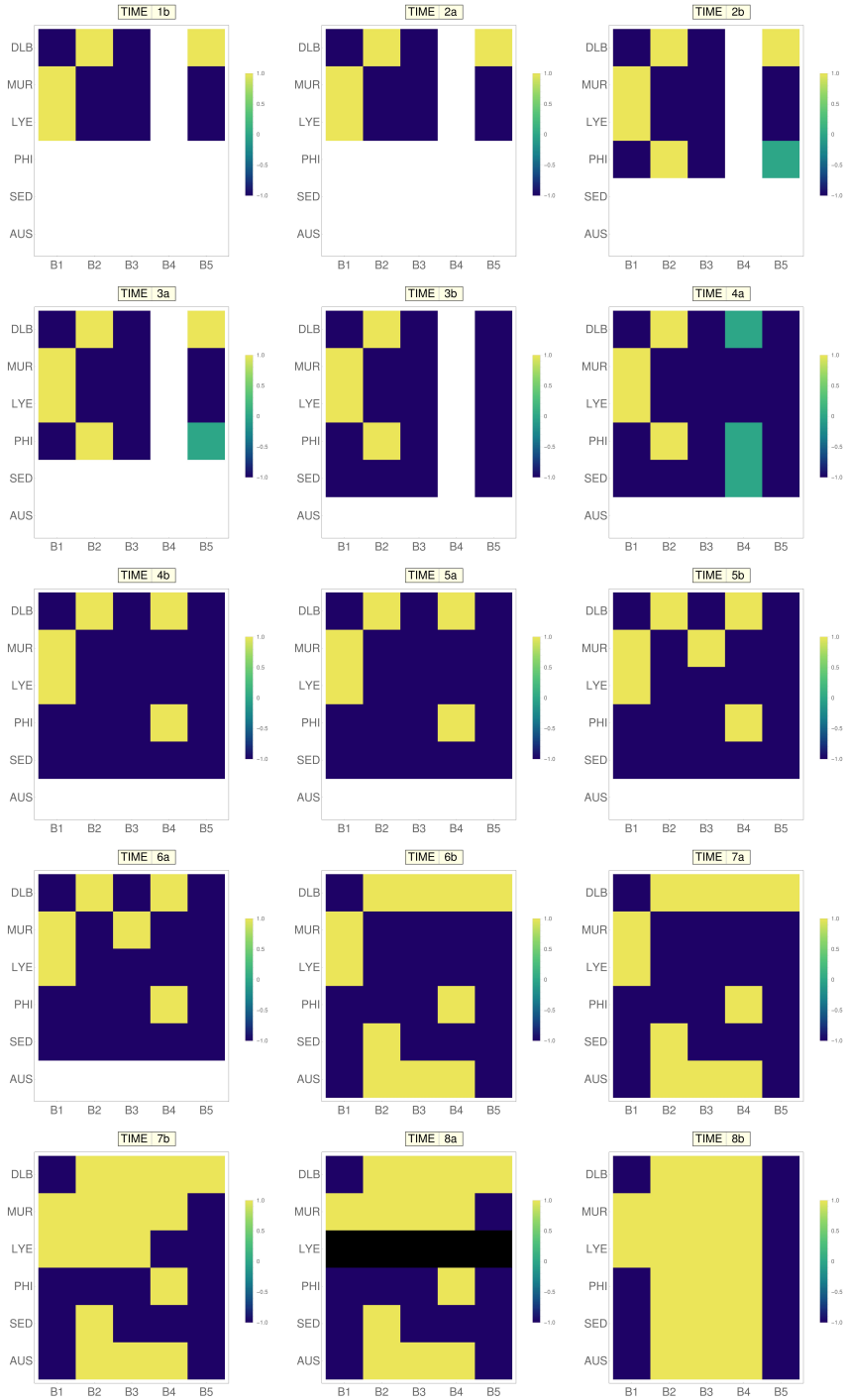

Figure 3.5.: Exogenously Defined Groups. A set of plots is shown with every plot corresponding to a certain time step. Every single plot shows the attitudes of the main participants towards a sentence $B1, ..., B5$, so-called *interpretative boundaries*. Possible attitudes towards these sentences are acceptation (yellow), judgment suspension (green) and rejection (blue). Not all sentences and persons are part of the debate right from the start. White spaces indicate their absence. A dialectically inconsistent position (black) is discarded from further analyses.

Time	Persons	B1–Groups	B2–Groups	B3–Groups	B4–Groups	B5–Groups
S1b	DLB, MUR, LYE	(MUR, LYE)	(MUR, LYE)	(DLB, MUR, LYE)		(MUR, LYE)
S2a	"	"	"	"		"
S2b	+ PHI	(MUR, LYE), (DLB, PHI)	(MUR, LYE), (DLB, PHI)	(DLB, MUR, LYE, PHI)		(MUR, LYE)
S3a	"	"	"	"		"
S3b	+ SED	(MUR, LYE), (DLB, PHI, SED)	(MUR, LYE, SED), (DLB, PHI)	(DLB, MUR, LYE, PHI, SED)		(DLB, MUR, LYE, PHI, SED)
S4a	"	"	"	"	(MUR, LYE), (DLB, PHI, SED)	"
S4b	"	"	(MUR, LYE, PHI, SED)	"	(MUR, LYE, SED), (DLB, PHI)	"
S5a	"	"	"	"	"	"
S5b	"	"	"	(DLB, LYE, PHI, SED)	"	"
S6a	"	"	"	"	"	"
S6b	+ AUS	(MUR, LYE), (DLB, PHI, SED, AUS)	(MUR, LYE, PHI), (DLB, SED, AUS)	(MUR, LYE, PHI, SED), (DLB, AUS)	(MUR, LYE, SED), (DLB, PHI, AUS)	(MUR, LYE, PHI, SED, AUS)
S7a	"	"	"	"	"	"
S7b	"	"	(DLB, MUR, LYE, SED, AUS)	(PHI, SED), (DLB, MUR, LYE, AUS)	(LYE, SED), (DLB, MUR, PHI, AUS)	"
S8a	"	(DLB, PHI, SED, AUS)	(DLB, MUR, SED, AUS)	(PHI, SED), (DLB, MUR, AUS)	(DLB, MUR, PHI, AUS)	(MUR, PHI, SED, AUS)
S8b	"	(MUR, LYE), (DLB, PHI, SED, AUS)	(DLB, MUR, LYE, PHI, SED, AUS)	(DLB, MUR, LYE, PHI, SED, AUS)	(DLB, MUR, LYE, PHI, SED, AUS)	(DLB, MUR, LYE, PHI, SED, AUS)

Figure 3.6.: Clustering persons according to *interpretative boundaries* $B1, ..., B5$. Some persons form a BX-group, if they exhibit the same attitude towards *the interpretative boundary* BX. For all persons, attitudes towards interpretative boundaries are shown in Fig. 3.5.

3.2.2. Endogenously Defined Groups

Groups are defined endogenously using different similarity thresholds. Similarity (SIM) is considered as a function with two arguments ranging from 0 to 1, see definition 3.2.2. Two dating hypotheses are more similar, if they share more atomic dating hypotheses.[5] Two bodies of evidence are more similar, if they share more evidential beliefs, that is, there are fewer contradictions or fewer judgment suspensions, compare example 3.2.1. To each similarity measure, there is a corresponding unsimilarity measure ($1 - SIM$). Comparing two dating hypotheses, the unsimilarity measure is a normalized Hamming distance. Comparing two bodies of evidence, the unsimilarity measure is - despite similarities - no normalized edit distance.[6]

Definition 3.2.2: Similarity

The degree of similarity between two dating hypotheses, $SIM(h_1, h_2)$, and the degree of similarity between two bodies of evidence, $SIM(e_1, e_2)$, are defined as follows.

$$\begin{aligned} SIM(h_1, h_2) &= 1 - \frac{|C|}{15} \\ SIM(e_1, e_2) &= 1 - \frac{-3*|C|-|S|}{-3*(|C|+|A|)-|S|} \end{aligned}$$

Here, for two sets of statements, $|C|$, $|S|$ and $|A|$ denote the number of contradictions, judgment suspensions and agreements, respectively.

[5] For some short analysis of differences between *theoretical distance* and the similarity between two dating hypotheses, see sec. C.1.

[6] For simplicity, presume no weights and no normalization, that is,

1. $1 - SIM(e_1, e_2) = |C| + |S|$

2. For two bodies of evidence e_1 and e_2 with $|e_2| > |e_1| : ED(e_1, e_2) \leq |e_2|$

Here, $ED(e_1, e_2)$ denotes the edit distance between e_1 and e_2. Consider the following two bodies of evidence, $e_1 = (1, 2, 3, 5, 6, 9, 11)$ and $e_2 = (\neg 1, \neg 3, \neg 5, \neg 6, \neg 9, 12, 13, 15, 20)$. It shows that $1 - SIM(e_1, e_2) > ED(e_1, e_2)$. Therefore, $1 - SIM(e_1, e_2)$ is no edit distance.

Example 3.2.1: Similarity Between Bodies of Evidence

Here, five bodies of evidence, namely $e_1 ... e_5$, are shown. For certain pairs, similarity is calculated.

$$
\begin{aligned}
e_1 &= (\neg 16, 17, 18, \neg 19, 20) \\
e_2 &= (\neg 16, 17, 18, \neg 19, 20) \\
e_3 &= (16, \neg 17, \neg 18, 19, \neg 20) \\
e_4 &= (16, \neg 17, \neg 18, 19, \neg 20, 26, \neg 27) \\
e_5 &= (21, 22, \neg 23, \neg 24, 25, 26, \neg 27) \\
1 &= SIM(e_1, e_2) \\
0 &= SIM(e_1, e_3) = SIM(e_1, e_4) = SIM(e_1, e_5)
\end{aligned}
$$

Similarity results are shown in Fig. 3.7. First, some words on similarity between two dating hypotheses. It ranges from 0.6 to 1. Hence, there are at most 6 contradictions. Except during the middle section, there are always two persons which maximally agree. However, they are not always the same. At the very end, everybody maximally agrees with one another. The two most unsimiliar dating hypotheses are those of De la Beche and Murchison respectively Lyell at time steps $7b$ and $8a$. At first, from time step $1b$ until $4a$, the similarity is either maximal or relatively low. Then, after time step $4a$, the similarity spectrum broadens. Diversity lasts until the penultimate time step.

Second, some words on similarity between two bodies of evidence. It ranges from 0.541 to 0.988. Hence, never, not even at the end, there are two maximally similar bodies of evidence. Hence, there are always either contradictions or judgment suspensions. The two most unsimiliar bodies of evidence are those of De la Beche and Murchison at time step $1b$. However, at the same time, those of Lyell and Murchison are not very similar either. At first, from time step $1b$ until $4a$, similarity increases. This is true for all pairs of persons. However, some pairs of persons are getting more similar than others. Then, after time step $4a$, similarity decreases for all pairs of persons. From time step $6b$ until the penultimate time step, similarity dynamics are not that simple, showing increase as well as decrease at the same time. At the last time step, similarity increases once again for all pairs of persons.

Third, some words on differences between dating hypotheses and bodies of evidence. It shows that two persons can, at the same time, accept remarkably similar bodies

of evidence but unsimilar dating hypotheses, and vice versa. At the beginning, at time steps $3b$ and $4a$, Sedgwick and De la Beche respectively Phillips accept rather similar bodies of evidence but unsimilar dating hypotheses. The same is true for De la Beche and Phillips from time step $4b$ to $6b$ and De la Beche and Austen from time step $6b$ until the penultimate one. As already mentioned, at time steps $1b$ and $2a$, Murchison and Lyell accept rather unsimiliar bodies of evidence while accepting the same dating hypothesis. The same is true, however not to the same degree of unsimilarity, for all time steps and two persons accepting the same dating hypothesis, compare for example all pairs of persons at the final step, or Phillips and Sedgwick from time step $4b$ to $6a$.

Fourth, some words on similarities between dating hypotheses and bodies of evidence. It shows that the average degree of similarity is maximal at the final step. This is true for dating hypotheses as well as for bodies of evidence. Finally, the average degree of similarity is 1.0 with regard to dating hypotheses and slightly less than 1.0 with regard to bodies of evidence. Hence, for the great Devonian controversy, dating hypotheses and bodies of evidence change such that they are finally much more similar, not at last due to argumentation. In so far as dating hypotheses and bodies of evidence together constitute a paradigm, this result may be understood as an illustration of Kuhn (1983) stating that scientific revolutions are not only triggered but also resolved by inter-paradigmatic exchange of arguments.[7]

Based on these similarity results, groups are defined endogenously using a similarity threshold, see definition 3.2.3 and example 3.2.2. Some persons form a group, if they are mutually sufficiently similar. Comparing five different similarity thresholds, it shows that similarity clusters depend on this threshold. In the following, results for two similarity thresholds will be presented in some more detail. See Appendix C for results relying on other thresholds. It shows that similarity clusters change with changes in the body of shared background beliefs, but not very often.

[7]See sec. 3.1.1 for some remarks on differences between paradigms and bodies of evidence.

Definition 3.2.3: Clustering Sufficiently Similar Persons

Be s_0 some similarity threshold. There is a similarity group G of n persons regarding this thresold, iff

$$\forall i, j \in G : SIM_{i,j} > s_0$$

Note that this translates into $\binom{n}{2} = \frac{n!}{2!(n-2)!}$ single equations.

Example 3.2.2: Clustering of Sufficiently Similar Bodies of Evidence

Here, s_0 is a similarity threshold and $1, ..., 6$ refer to persons. The following table spells out definition 3.2.3.

Group members	Conditions	
$1, 2$		$SIM(e_1, e_2) > s_0$
$1, 2, 3$	$1, 2$ form a group and	$SIM(e_1, e_3) > s_0$, $SIM(e_2, e_3) > s_0$
$1, 2, 3, 4$	$1, 2, 3$ form a group and	$SIM(e_1, e_4) > s_0$, $SIM(e_2, e_4) > s_0$, $SIM(e_3, e_4) > s_0$
$1, 2, 3, 4, 5$	$1, 2, 3, 4$ form a group and	$SIM(e_1, e_5) > s_0$, $SIM(e_2, e_5) > s_0$, $SIM(e_3, e_5) > s_0$, $SIM(e_4, e_5) > s_0$
$1, 2, 3, 4, 5, 6$	$1, 2, 3, 4, 5$ form a group and	$SIM(e_1, e_6) > s_0$, $SIM(e_2, e_6) > s_0$, $SIM(e_3, e_6) > s_0$, $SIM(e_4, e_6) > s_0$, $SIM(e_5, e_6) > s_0$

Fig. 3.10 shows clustering results for a similarity threshold of 1.0. Considering bodies of evidence, there are no groups at all. However, considering dating hypotheses, there are groups and they are the same as those in Fig. 3.2 which is taken from (Rudwick, 1988). First, there is only one group, namely Murchison and Lyell. With Phillips entering the stage, there is a second one, namely Phillips and De la Beche. Entering the stage, Sedgwick first joins the group of Murchison and Lyell. At time

step 4*b*, leaving their former groups, Sedgwick and Phillips form a new group which is the only group left after the next individual belief change. At time step 6*b*, the old group is replaced by a new one consisting of Murchison, Lyell and Phillips. However, Phillips leaves this group two steps before the final consensus.

Fig. 3.11 shows clustering results for a similarity threshold of 0.85. Considering dating hypotheses, first, results are the same as for a similarity threshold of 1.0. At time step 4*b*, there is a single group consisting of all persons but De la Beche, which, after the next individual belief change, is left by Murchison being no longer sufficiently similar with Phillips and Sedgwick. At time step 6*b*, there are two groups, namely the one known from time step 4*b* and a new one consisting of Murchison, Lyell, Phillips and Austen. However, as soon as time step 7*b*, Phillips and Sedgwick leave their groups being no longer sufficiently similar with Murchison and Lyell. Considering bodies of evidence, first, results are the same as for a similarity threshold of 1.0, that is, there is no group. However, with Phillips entering the stage, there is a group, namely Phillips and De la Beche. Entering the stage, Sedgwick joins this group and stays there for quite a while. At time step 6*b*, Sedgwick splits, being no longer sufficiently similar with De la Beche. At the same time, entering the stage, Austen is sufficiently similar with all persons but Murchison and Lyell which form a group of their own. At time step 7*a*, with Sedgwick being sufficiently similar to De la Beche again, there is a group which is the same as those known from 4*b* extended by Austen. However, after changing their individual beliefs, Phillips and Sedgwick stop being sufficiently similar to De la Beche. Finally, all persons are mutually sufficiently similar.

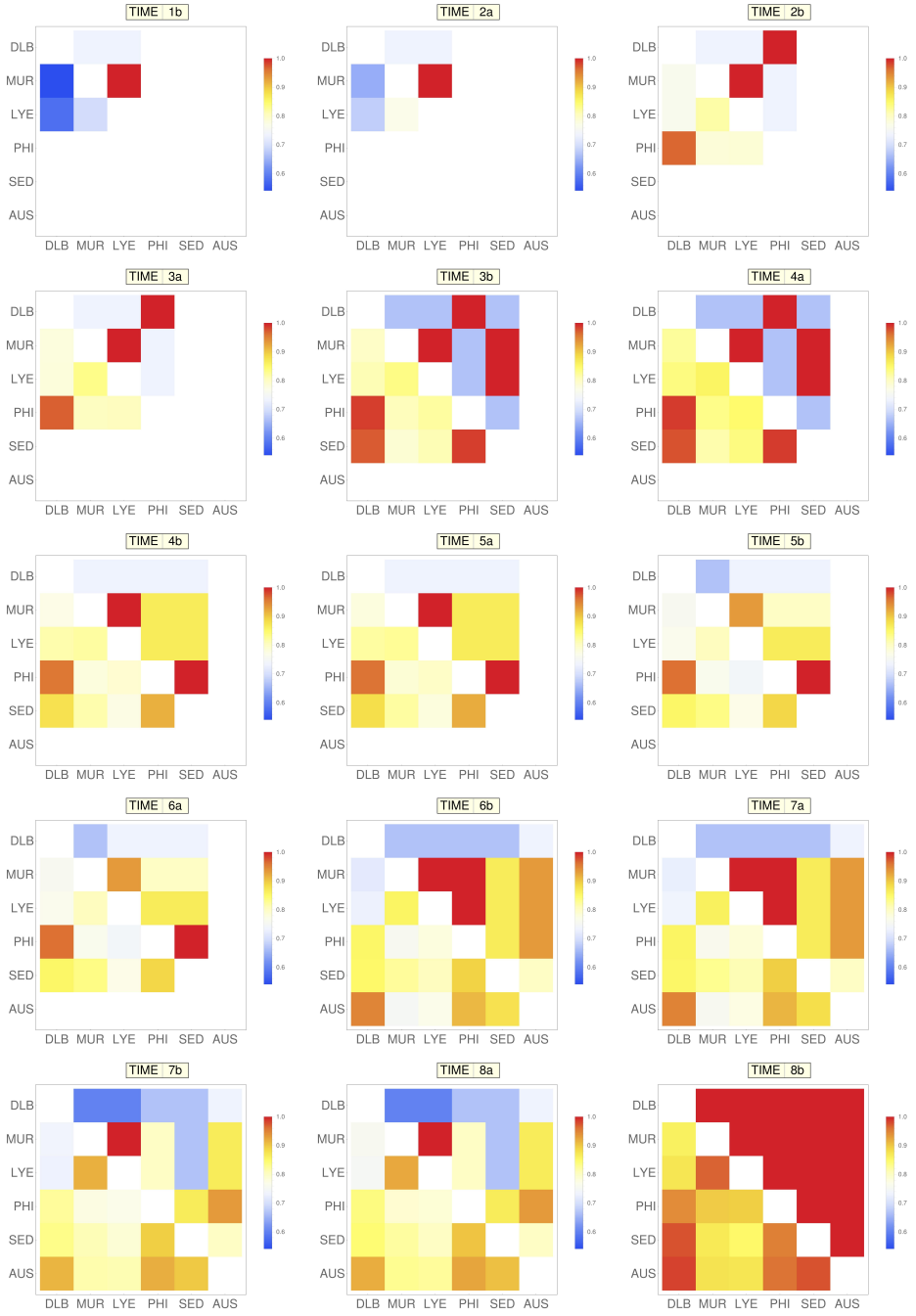

Figure 3.7.: A set of plots is shown with every plot corresponding to a certain time step t. Every single plot consists of an upper and a lower triangle showing similarity between two dating hypotheses and bodies of evidence, respectively. These hypotheses and bodies of evidence belong to De la Beche (DLB), Murchison (MUR), Lyell (LYE), Sedgwick (SED) and Austen (AUS). Similarity ranges from 0.54 to 1 and is represented using a color function, turning from blue over yellow to red.

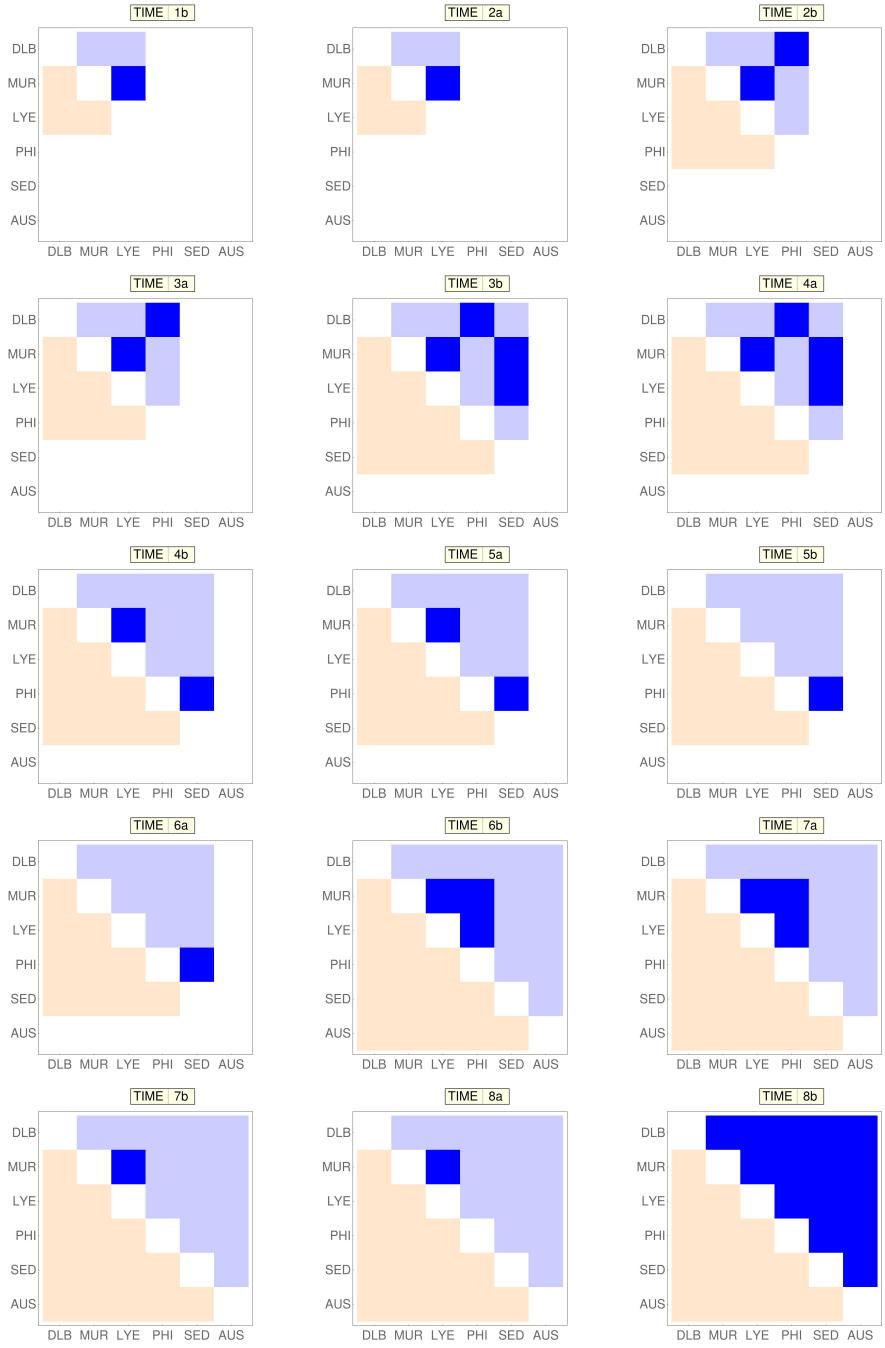

Figure 3.8.: A set of plots is shown with every plot corresponding to a certain time step. Every single plot consists of an upper and a lower triangle showing similarity between two dating hypotheses and bodies of evidence, respectively. These hypotheses and bodies of evidence belong to De la Beche (DLB), Murchison (MUR), Lyell (LYE), Sedgwick (SED) and Austen (AUS). Darker and lighter shading represent a similarity of 1 and less than 1, respectively.

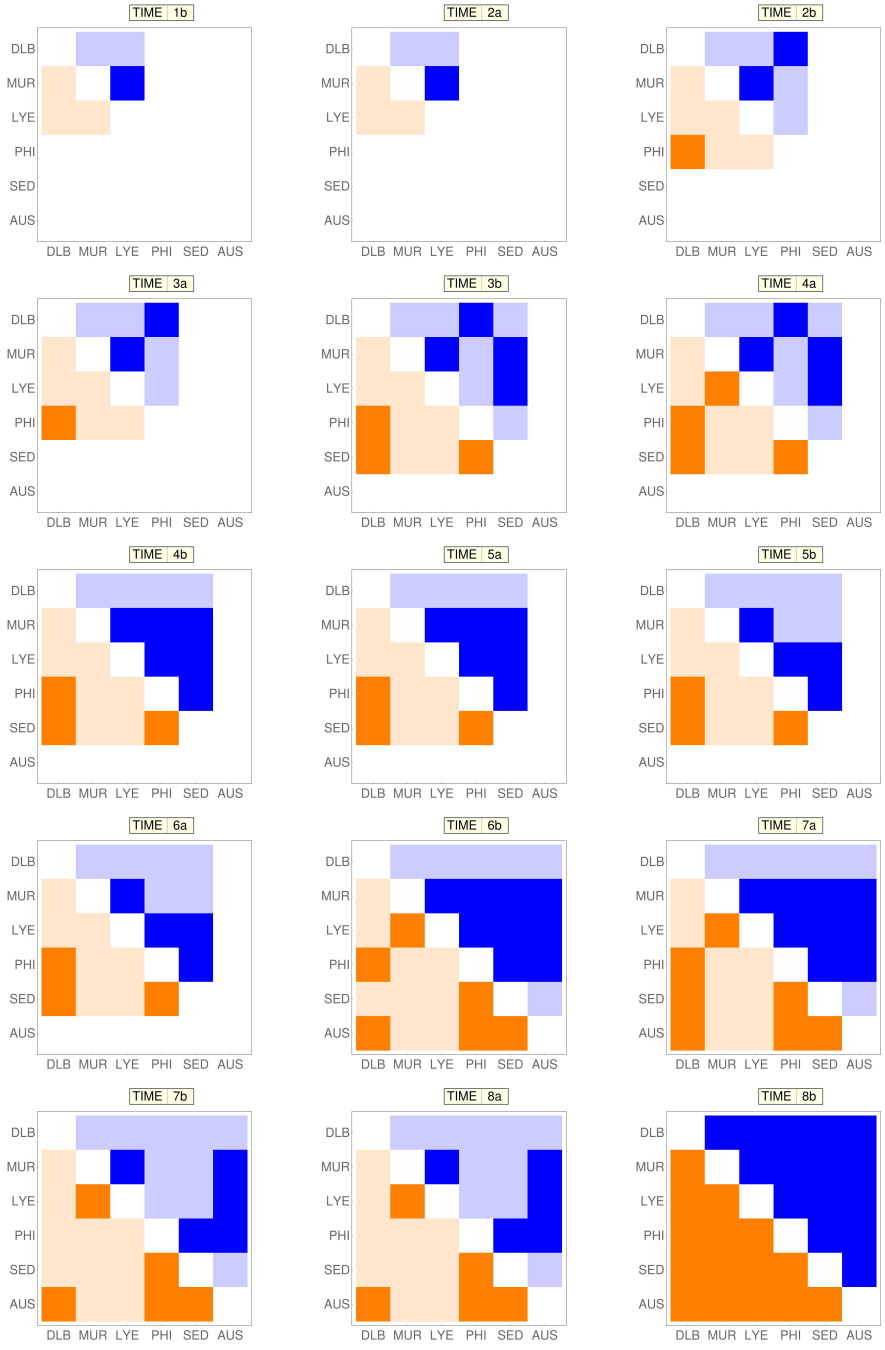

Figure 3.9.: A set of plots is shown with every plot corresponding to a certain time step. Every single plot consists of an upper and a lower triangle showing similarity between two dating hypotheses and bodies of evidence, respectively. These hypotheses and bodies of evidence belong to De la Beche (DLB), Murchison (MUR), Lyell (LYE), Sedgwick (SED) and Austen (AUS). Darker and lighter shading represent a similarity greater or equal 0.85 and less than 0.85, respectively.

SIM≥1.0			
Time	Persons	H–Groups	E–Groups
S1b	DLB, MUR, LYE	(MUR, LYE)	–
S2a	"	"	"
S2b	+ PHI	(MUR, LYE), (DLB, PHI)	"
S3a	"	"	"
S3b	+ SED	(MUR, LYE, SED), (DLB, PHI)	"
S4a	"	"	"
S4b	"	(MUR, LYE), (PHI, SED)	"
S5a	"	"	"
S5b	"	(PHI, SED)	"
S6a	"	"	"
S6b	+ AUS	(MUR, LYE, PHI)	"
S7a	"	"	"
S7b	"	(MUR, LYE)	"
S8a	"	"	"
S8b	"	(DLB, MUR, LYE, PHI, SED, AUS)	"

Figure 3.10.: Clustering persons according to a similarity threshold of 1. Clustering is performed based on similarities between dating hypotheses (H-groups) and bodies of evidence (E-groups), compare also Fig. 3.8

SIM≥0.85			
Time	Persons	H–Groups	E–Groups
S1b	DLB, MUR, LYE	(MUR, LYE)	–
S2a	"	"	"
S2b	+ PHI	(MUR, LYE), (DLB, PHI)	(DLB, PHI)
S3a	"	"	"
S3b	+ SED	(MUR, LYE, SED), (DLB, PHI)	(DLB, PHI, SED)
S4a	"	"	(DLB, PHI, SED), (MUR, LYE)
S4b	"	(MUR, LYE, PHI, SED)	(DLB, PHI, SED)
S5a	"	"	"
S5b	"	(LYE, PHI, SED), (MUR, LYE)	"
S6a	"	"	"
S6b	+ AUS	(MUR, LYE, PHI, AUS), (MUR, LYE, PHI, SED)	(MUR, LYE), (PHI, SED, AUS), (DLB, PHI, AUS)
S7a	"	"	(MUR, LYE), (DLB, PHI, SED, AUS)
S7b	"	(MUR, LYE, AUS), (PHI, AUS), (PHI, SED)	(MUR, LYE), (DLB, AUS), (PHI, SED, AUS)
S8a	"	"	"
S8b	"	(DLB, MUR, LYE, PHI, SED, AUS)	(DLB, MUR, LYE, PHI, SED, AUS)

Figure 3.11.: Clustering persons according to a similarity threshold of 0.85. Clustering is performed based on similarities between dating hypotheses (H-groups) and bodies of evidence (E-groups), compare also Fig. 3.9

3.2.3. Belief Changes

How do scientists change their beliefs individually? Do they only shift between dialectically consistent positions? Are there only belief changes regarding hypotheses? Or do scientists hold on to certain hypotheses while changing evidential beliefs?

According to Popper (1935), a belief change is always triggered by some new evidence, e, falsifying some hypothesis, h, that is, e is contradictory to some deductive implication of h. Lakatos (1970) refines Popper (1935) by introducing hard-core and auxiliary assumptions. According to Lakatos (1970), there are always hypotheses which are given up only very reluctantly. In order to protect these so-called hard-core assumptions against falsification, other assumptions which are only presumed for auxiliary reasons are given up. Are belief systems revised as a whole or rather in a piecemeal and reluctant way? According to Laudan (1984, p. 74), the latter holds: "[...] the scientist will have compelling reasons for replacing one component or other of his world view with an element that does the job better. Yet he need not modify everything else."

How do participants of the great Devonian controversy change their beliefs individually? How do they shift from one group of a dating hypothesis and a body of evidence to another one?

First, is there only shifting between dialectically consistent positions? All main participants possess a dialectically consistent position all of the time, the only exception being Lyell at time step $8a$. At this time step, Lyell assigns complementary truth values to a certain sentence, namely a statement about Scottish Old Red Sandstone fossils. As already noted in sec. 3.1.3, moving from time step $7b$ to $8a$, there is a new dialectical structure and there are some new *shared background beliefs*. However, Lyell's individual beliefs stay the same. At time step $7b$, Lyell infers from his body of evidence and the dialectical structure that the Non-Culm strata are Old Red Sandstone in age. At $8a$, based on some new *shared background beliefs* and arguments, he accepts a new definition of some characteristics of the Old Red Sandstone period. Based on another new *shared background belief* and argument, he finally infers a statement about Scottish Old Res Sandstone fossils which he rejects, at the same time, as a consequence of his principle. Therefore, Lyell's position at time step $8a$ is dialectically inconsistent. Lyell solves this problem by shifting to a limited version of his principle, allowing Scottish Old Red Sandstone fossils not being intermediate in character between those of Silurian and Mountain Limestone

strata.

Second, are there only belief changes regarding dating hypotheses? It shows that participants of the great Devonian controversy do not only change their dating hypotheses, but also their evidential beliefs. Given that participants are rational, this result dis-confirms strict falsificationism in the sense of Popper (1935). However, note that Popper (1935) presumes that (*i*) a hypothesis is a universal statement and (*ii*) evidential beliefs are singular and existential. In this thesis, a dating hypothesis is no universal statement and evidential beliefs are not always singular and existential. Dating hypotheses and evidential beliefs are not always changed at the same time. Often, participants hold on to a certain dating hypothesis while changing evidential beliefs. Hence, for the great Devonian controversy, often, dating hypotheses and evidential beliefs can be considered as hard core and auxiliary assumptions in the sense of Lakatos (1970), respectively.

Third, are dating hypotheses and bodies of evidence revised as a whole? Most of the time, there are only minor relative changes in similarity. Hence, most of the time, dating hypotheses as well as bodies of evidence are only slightly altered. Therefore, beliefs are not revised as a whole, but rather in a piecemeal and reluctant way. This result illustrates (Laudan, 1984). However, note that world views as introduced in (Laudan, 1984) comprise some statement types not figuring in my reconstruction, compare sec. 3.1.1.

Fourth, are there some atomic dating hypotheses or evidential beliefs which are kept come what may? For the great Devonian controversy, there are some atomic dating hypotheses and evidential beliefs which are constantly kept, or at least very reluctantly given up. Hence, these atomic dating hypotheses or evidential beliefs can be considered as hard core assumptions in the sense of Lakatos (1970). Take for example Murchison holding on to his denial of Carboniferous fossils in strata older than Old Red Sandstone and dating the main part of the Culm strata as Coal Measures. As examples for beliefs which are rejected only very reluctantly consider De la Beche holding on to date the Non-Culm strata as Cambrian, or Lyell holding on to his principle. However, there are also some atomic dating hypotheses and evidential beliefs which are revised several times. Take as an example De la Beche returning to his initial belief that the main part of the Culm strata is in the middle of the sequence before giving it up at the final time step. Take as another example Murchison and its attitudes towards the assumption of some Non-Culm strata being

Old Red Sandstone in age.

What reasons are there for belief revisions? Answering this question is not in the scope of this subsection, but in the scope of this thesis. The next section analyses belief changes, however, only in terms of confirmation.

3.3. Confirmation Dynamics

In the following chapter, confirmation dynamics are presented. Three different notions of confirmation are compared, namely $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. These are the same as those analyzed in the first part of my thesis, that is chapter 2.[8] The main results are as follows:

- ($H1.4$) The great Devonian controversy starts and ends with all main participants accepting a dating hypothesis with a maximal degree of confirmation (given a certain evidence), that is with a degree of 1, independent of a certain confirmation measure.
- ($H1.5$) For most time steps and participants, $DOJ(h|e)$ and $Z_{DOJ}(h,e)$ are rather similar, both in value and relative changes, and much smaller than 1.
- ($H1.6$) For most time steps and participants, $DOJ(h|e)$ and $F_{DOJ}(h,e)$ are rather unsimilar, both in value and relative changes. For most time steps and participants, $F_{DOJ}(h,e)$ increases with increasing $\frac{DOJ(h|e)}{DOJ(h)}$ and is fairly approximated by 1.

The last part of this chapter is about rational belief change. Is rationality in belief change related with some kind of evidential support? Here, evidential support is spelled out in terms of $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. The following principles of rational belief change are tested:

- ($RAT1$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it does not decrease the dating hypothesis's degree of evidential support.

[8] Note that, for the great Devonian controversy, analyzing belief changes with regard to dating hypotheses in terms of confirmation, some confirmation theories are not applicable due to dating hypotheses being no general statements. See for example hypothetico-deductivism as described in (Schurz, 2006).

- ($RAT2$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it maximizes the dating hypothesis's degree of evidential support.

Here, a person maximizes her dating hypothesis's evidential support, if she chooses a dating hypothesis such that, given her body of evidence, there is no better confirmed dating hypothesis.

For the great Devonian controversy, the following shows:

- ($H4.1$) Most of the time, individual belief changes are rational. However, there are individual belief changes which are not rational. Using $F_{DOJ}(h, e)$, individual belief changes are *more* often rational than using one of the other two confirmation measures.
- ($H4.2$) Shared belief changes are *less* often rational than individual belief changes. Using $F_{DOJ}(h, e)$, shared belief changes are *more* often rational than using one of the other two confirmation measures.

Presuppose that participants of the great Devonian controversy are rational. Together with $H4.1$, it follows that there are exceptions to the two previously introduced principles of rational belief change. This confirms Feyerabend (1976) stating that there is no scientific rule without any exceptions. However, it does not support relativism in the sense of "anything goes".

3.3.1. Comparing Absolute and Relevance Confirmation

First, for every time step and main participant, the dating hypothesis's degree of justification, $DOJ(h|e)$, is calculated, see Fig. 3.12. It shows that, for most time steps and participants, $DOJ(h|e)$ is much smaller than one.

$$DOJ(h|e) \ll 1 \tag{3.1}$$

For all participants at time steps $S0$ and $S8$, it holds: $DOJ(h|e) = 1$. Hence, the great Devonian controversy starts and ends with all participants accepting a dating hypothesis which is implied by her body of evidence (and the dialectical structure). However, for most time steps and participants, it holds: $DOJ(h|e) < 1$, that is,

the dating hypothesis is not implied by the body of evidence (and the dialectical structure). Most of the time, only some atomic dating hypotheses are implied by the body of evidence (and the dialectical structure). For example, at $S1$, based on his body of evidence (and the dialectical structure), Murchison infers some dating for the main part of the Culm strata, but no dating for the black Culm limestone or the Non-Culm strata. Note further, that a certain dating of some strata is implied only by a compound of sentences, compare sec. 3.1.1. Very often, only some but not all of these sentences are accepted.

DLB	((0b, 1.), (1a, 1.), (1b, 1.), (2a, 1.), (8b, 1.))
MUR	((7b, 1.), (8a, 1.), (8b, 1.))
LYE	((7b, 1.), (8b, 1.))
PHI	((8b, 1.))
SED	((8b, 1.))
AUS	((8b, 1.))

Table 3.1.: Participants and time steps for which it holds: The dating hypothesis is deductively inferred from the body of evidence and dialectical structure.

Second, for every time step and main participant, $Z_{DOJ}(h, e)$ is calculated. It shows that, for most time steps and participants, $DOJ(h|e)$ and $Z_{DOJ}(h, e)$ are rather similar, both in value and relative changes, see Fig. 3.13.

$$Z_{DOJ}(h, e) \approx DOJ(h|e) \tag{3.2}$$

This can be explained by recapitulating definition 2.1.8 and using some additional results, namely Fig. D.4 and Fig. D.3, showing that the following equations hold.

$$DOJ(h) \ll 1 \tag{3.3}$$

$$\frac{DOJ(h|e)}{DOJ(h)} \gg 1 \tag{3.4}$$

Third, $F_{DOJ}(h, e)$ is calculated. It shows that, for most time steps and participants,

$DOJ(h|e)$ and $F_{DOJ}(h,e)$ are rather unsimilar, both in value and relative changes, compare Fig. 3.14 with Fig. 3.12. For most time steps and participants, $F_{DOJ}(h,e)$ increases with increasing $\frac{DOJ(h|e)}{DOJ(h)}$ and is fairly approximated by 1. The following approximation of $F_{DOJ}(h,e)$ meets the first condition independent of a certain ratio of $\frac{DOJ(h|e)}{DOJ(h)}$ and the second condition, if equation 3.4 holds.

$$F_{DOJ}(h,e) \sim \frac{\frac{DOJ(h|e)}{DOJ(h)} - 1}{\frac{DOJ(h|e)}{DOJ(h)} + 1} \tag{3.5}$$

This approximation of $F_{DOJ}(h,e)$ can be explained by recapitulating definition 2.1.7, using equations 3.1, 3.3 and 3.4 as well as an additional result, namely Fig. D.2, showing that the following inequality holds.

$$DOJ(e) \neq 0 \tag{3.6}$$

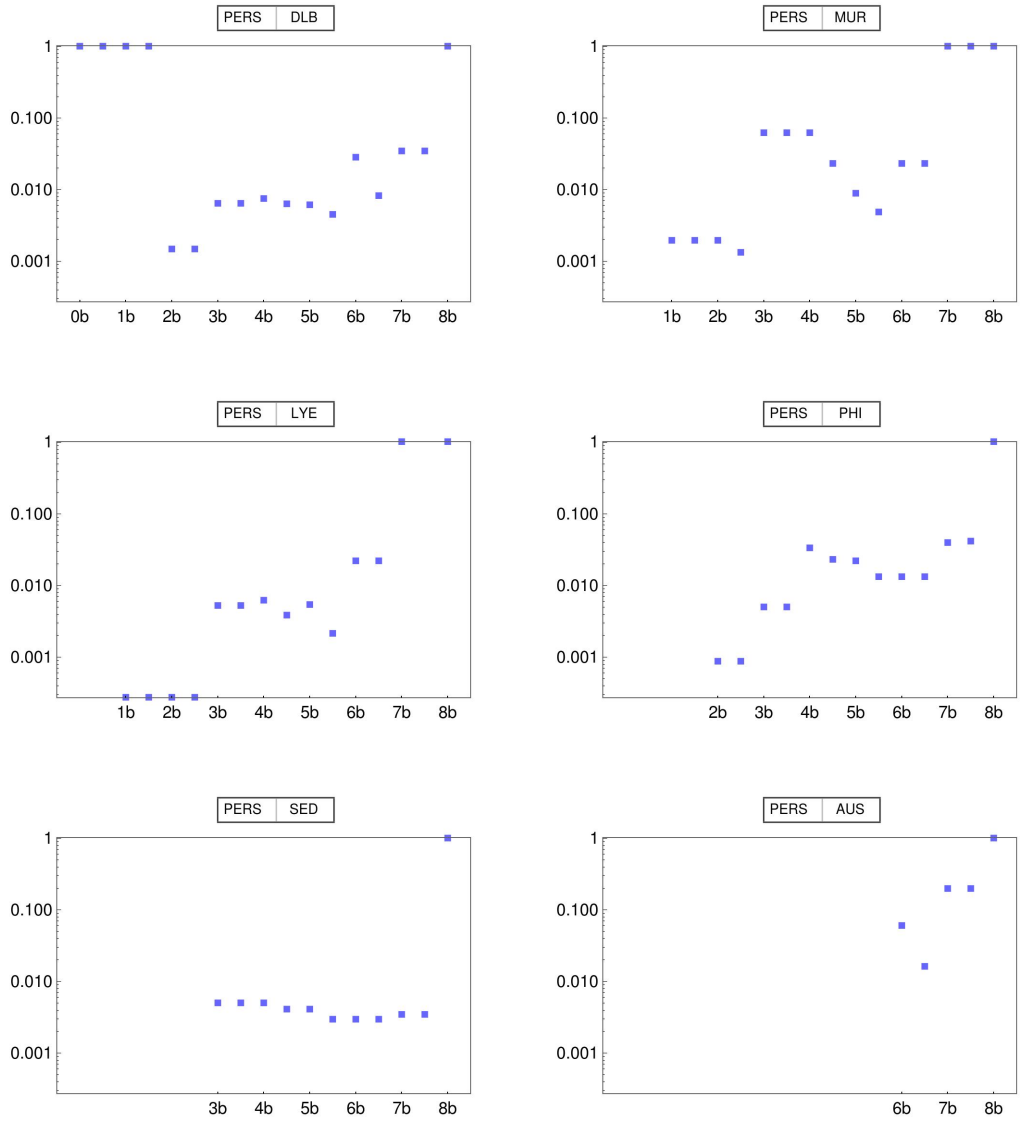

Figure 3.12.: Confirmation Dynamics using $DOJ(h|e)$. $DOJ(h|e)$ ranges from 0.00028 to 1.

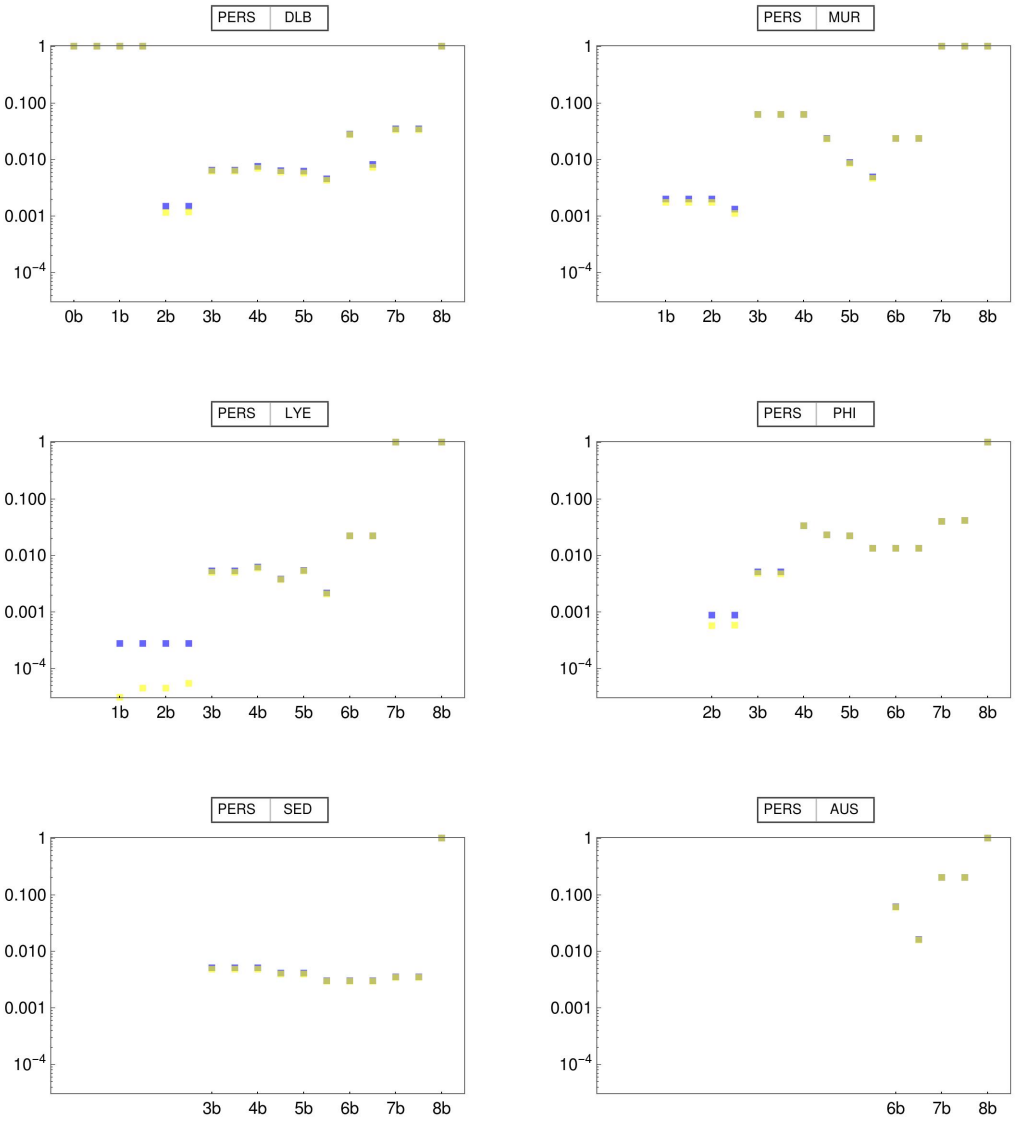

Figure 3.13.: Comparison of $Z_{DOJ}(h, e)$ and $DOJ(h|e)$. Yellow and blue represent $Z_{DOJ}(h, e)$ and $DOJ(h|e)$, respectively. $Z_{DOJ}(h, e)$ ranges from 0.00003 to 1.

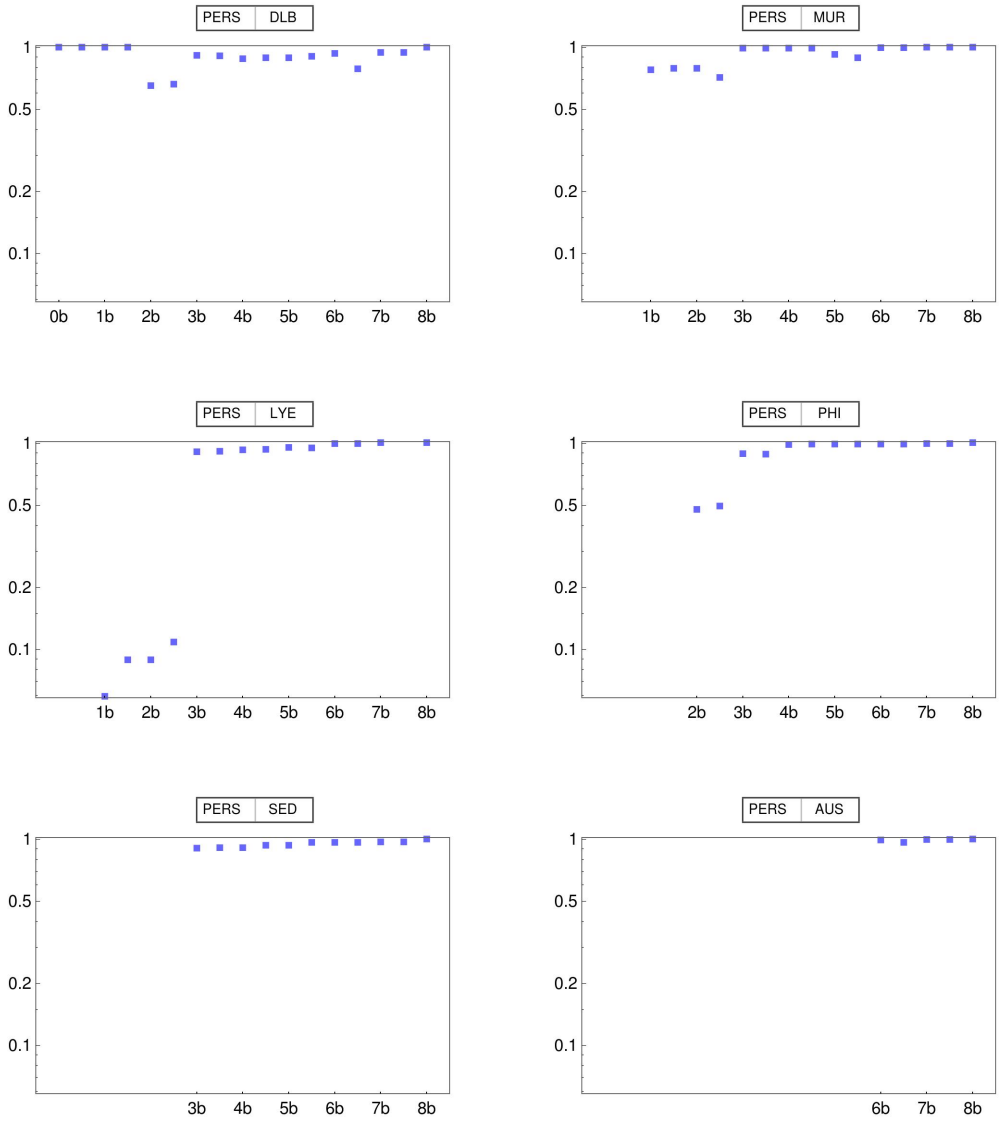

Figure 3.14.: Confirmation Dynamics using $F_{DOJ}(h, e)$. $F_{DOJ}(h, e)$ ranges from 0.05931 to 1.

3.3.2. Belief Changes and Evidential Support

Among philosophers of science, opinions differ on what a rational belief change looks like and what makes it rational. Is rationality in belief change related with some kind of evidential support?[9] Here, evidential support is understood as a type of evidential justification and spelled out in terms of three different Bayesian confirmation measures, compare the previous subsection.

As suggested by Popper (1979), does a rational agent try to falsify his dating hypothesis, that is, does he try to find some evidence implying the negation of his dating hypothesis? Or does he verify his dating hypotheses, that is, he accepts some evidence implying his dating hypothesis? Or is there some other principle of rational belief change? Consider as promising candidates the two following ones:

- ($RAT1$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it does not decrease the dating hypothesis's degree of evidential support.

- ($RAT2$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it maximizes the dating hypothesis's degree of evidential support.

Here, a person maximizes her dating hypothesis's evidential support, if she chooses a dating hypothesis such that, given her body of evidence, there is no better confirmed dating hypothesis, see definition 3.3.1.

[9] Certainly, there is more to think about. Take as examples belief changes from one belief system to another, which promote fruitfulness, simplicity or progressiveness, see for example (Kuhn, 1983), (Lakatos, 1970) and (Laudan, 1977)

Definition 3.3.1: Increasing and Maximizing Evidential Support

Consider a person j shifting beliefs from time step t_1 to t_2, that is shifting from $(h^j_{t_1}, e^j_{t_1})$ to $(h^j_{t_2}, e^j_{t_2})$. Evidential support increases, iff

$$CONF_{DOJ}(h^j_{t_2}, e^j_{t_2}) > CONF_{DOJ}(h^j_{t_1}, e^j_{t_1})$$

Evidential support is maximized, iff for all possible dating hypotheses h_k, it holds:

$$CONF_{DOJ}(h^j_{t_2}, e^j_{t_2}) \geq CONF_{DOJ}(h^k_{t_2}, e^j_{t_2})$$

Here, $CONF_{DOJ}$ is some confirmation measure relying on degrees of justification.

For the great Devonian controversy, are participants rational in the sense of the epistemic rule introduced in sec. 2.2? To investigate a participant's rationality in this sense at a certain time, several steps have to be performed. First, the situation has to be characterized in terms of higher-order evidence, namely the relative size and accuracy of the body of evidence and the inferential density of the dialectical structure. As a result of the first part of my thesis, for every situation characterized this way, there is a critical region of a statistical test with confirmation as test statistic. Second, a participant's degree of confirmation has to be compared with the critical region. Does it fall into this region?

This thesis does not answer the last question for several reasons. First, for both parts of my thesis, that is chapter 2 and chapter 3, the notion of accuracy of a body of evidence differs. For part one and two of my thesis, accuracy depends on the similarity with the truth and final consensus position, respectively. In the first part of my thesis, there is a truth, that is a position assigning a truth value to *all* sentences of a debate. Therefore, similarity with the truth is adequately captured by the amount of true evidence claims. A body of evidence is more similar to the truth, if the amount of true evidence claims is larger, that is, the amount of false evidential claims is smaller. Hence, accuracy is well defined by the ratio of true first-order evidence claims. For the great Devonian controversy, there is no truth, only a final consensus position. This position does *not* assign a truth value to *every* sentence of the debate. Therefore, similarity with the final consensus position has to account for judgment suspensions. A body of evidence is more similar to the final consensus

position, if they share more beliefs, that is, there are fewer contradictions *or* fewer judgment suspensions. Hence, the accuracy of a participant's body of evidence has to be defined otherwise than the ratio of beliefs shared with the final consensus, accounting for judgment suspensions. Second, values of higher-order evidence differ for both parts of my thesis. For example, for the great Devonian controversy, there is no time step, where values of the inferential density correspond to some used in the first part of my thesis, compare sec. 3.1.3 and sec. 2.2.1.

Let us take a closer look at the the great Devonian controversy and see if its participants form their beliefs rationally and according to confirmation. First, do participants falsify or verify their dating hypotheses? The previous subsection shows that, for all participants, it holds that the conditional degree of justification of her dating hypothesis is always greater 0, that is, her body of evidence never falsifies her dating hypothesis.[10] Further, only for a few time steps and participants, it holds that the conditional degree of justification of her dating hypothesis is 1, that is, her body of evidence verifies her dating hypothesis. Nevertheless, there is always a change in evidential support.

	DOJ	Z	F
Individual belief changes	$\frac{4}{41}$	$\frac{4}{41}$	$\frac{3}{41}$
Shared belief changes	$\frac{13}{41}$	$\frac{13}{41}$	$\frac{4}{41}$

Table 3.2.: Proportion of belief changes with decreasing evidential support.

Second, do participants of the great Devonian controversy always increase their dating hypotheses's evidential support? My analyses show that for some time steps and persons, shifting from one position to another, the degree of evidential support decreases, see Fig. 3.15. Changing shared background beliefs and the dialectical structure more often leads to a decrease in evidential support than changing individual beliefs, compare Tab. 3.2. This is true for all confirmation measures. However, this is *less* often the case using $F_{DOJ}(h, e)$.[11] Most of the time, individual belief changes lead to an increase in evidential support. This is true for all confirmation measures. However, this is *most* often the case using $F_{DOJ}(h, e)$.

What follows from there being decreases in evidential support? Consider the follow-

[10] Note that this doesn't exclude cases where there are two persons i and j such that i's body of evidence falsifies j's dating hypothesis.

[11] Recapitulate definitions and approximations of the three confirmations measure as given in sec. 2.1.2 and sec. D.1.

ing argument, mimicking one from (Laudan, 1990).

Example 3.3.1: Argument with Quineian Underdetermination

(QUDN) Any dating hypothesis can be rationally reconciled with any recalcitrant evidence by making suitable adjustments in our evidential beliefs.
(RAT1) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it does not decrease the dating hypothesis's degree of evidential support.

—

(EGAL) Every dating hypothesis is as well supported by the evidence as one of its rivals.

Note that $P1$ is a normative reading of Quineian Underdetermination and $EGAL$ states some egalitarian hypothesis. Based on my analyses, either $RAT1$ has to be rejected or some participants of the great Devonian controversy have to be considered as irrational, at least sometimes. Further, my analyses show that, for every confirmation measure, time step and person, there are at least two dating hypothesis differing in evidential support, see Fig. D.5, Fig. D.6 and Fig. D.7.[12] Hence, based on my analyses, $EGAL$ has to be rejected. Therefore, $RAT1$ and $QUDN$ cannot be accepted both at the same time.

What causes decreases in evidential support? New beliefs? Old beliefs, which are kept, come what may? In the following, I want to shortly delineate individual belief changes leading to decrease in evidential support. In the beginning, De la Beche infers dating hypotheses from the characteristic rock type principle. At time step $2b$, accepting local variations in sedimentation, De la Beche has to reject this criterion. As a consequence, De la Beche's dating hypothesis is no longer implied by his body of evidence and loses much of its evidential support. This is true for all three confirmation measures. Moving from time step $4a$ to $4b$, De la Beche faces a minor decrease in evidential support. However, this is negligible, if not using $F_{DOJ}(h, e)$. At time step $4b$, De la Beche slightly changes his dating hypothesis, accepting that some Non-Culm strata are Silurian in age. Still rejecting the characteristic fossils principle, this dating is not implied by his body of evidence. However, De la Beche's new dating hypothesis exhibits a higher degree of absolute justification. Moving

[12] Actually, these plots show a lot more, namely the whole spectrum of evidential support for a given body of evidence and a certain confirmation measure. Note that results depend heavily on the confirmation measure.

from time step $5a$ to $5b$, De la Beche, Phillips and Murchison face a decrease in evidential support. For the first two, there is a negligible decrease using $F_{DOJ}(h, e)$, and a minor one using the other two confirmation measures. For Murchison, there is a minor one using $F_{DOJ}(h, e)$ and a major decrease using the other two confirmation measures. What is new at time step $5a$? There are some new fossil findings, namely Carboniferous fossils in some north Devonian Non-Culm strata. These findings trigger some new arguments inferring that these north Devonian Non-Culm strata are Old Red Sandstone in age and being based on different criteria and auxiliary assumptions. At this very time step, given their evidential beliefs and the new dialectical structure, De la Beche and Phillips can infer some Non-Culm strata as being Old Red Sandstone in age by accepting some of these auxiliary assumptions as well as these fossil findings. Given their evidential beliefs, rejecting these fossil findings, all of these arguments lose their restrictive effects. However, at $5b$, De la Beche and Phillips both accept these fossil findings. In order to prevent themselves from being forced to change their dating hypotheses, they both deny one of the auxiliary assumptions, namely the amount of collected fossils being sufficiently large to draw inferences upon. Now, let us turn to Murchison. At time step $5b$, he also accepts these new fossil findings. Holding on to his denial of there being carboniferous fossils in strata older than Old Red Sandstone, Murchison is forced to date the north Devonian Non-Culm strata at least as young as Old Red Sandstone. As a consequence, he has to give up the characteristic fossil principle. This way, dating the main part of the Culm as Coal Measures is no longer implied by his body of evidence. This figures as a loss in evidential support.

	DOJ	Z	F
Individual belief changes	$\frac{13}{41}$	$\frac{17}{41}$	$\frac{8}{41}$
Shared belief changes	$\frac{8}{17}$	$\frac{19}{34}$	$\frac{9}{34}$

Table 3.3.: Proportion of belief changes where evidential support is not maximized.

Third, do participants of the great Devonian controversy maximize their dating hypotheses's evidential support? My analyses show that for some time steps and persons, shifting from one position to another, the degree of evidential support is not maximized, see Fig. 3.16. Changing shared background beliefs and the dialectical structure less often maximize evidential support than changing individual beliefs, compare Tab. 3.3. This is true for all confirmation measures. However, this is

less often the case using $F_{DOJ}(h, e)$.[13] Most of the time, individual belief changes maximize evidential support. This is true for all belief changes and confirmation measures, except shared belief changes and $Z_{DOJ}(h, e)$. However, this is *more* often the case using $F_{DOJ}(h, e)$.

Differences between maximal and actual evidential support are sometimes of different orders using different confirmation measures, see Fig. 3.16. Only if there is no such difference for absolute confirmation, then differences between maximal and actual evidential support are of the same order for the two relevance confirmation measures.

What reasons are there for not maximizing evidential support? Are there beliefs which are kept come what may? Are there personal aims which are pursued? Are there some persons considered as experts by some others? In the following, I want to shortly delineate individual belief changes not maximizing evidential support.[14]

Entering the stage, Murchison dates the main part of the Culm strata as Coal Measures, based on a certain fossil criterion and some fossil findings. As everybody else, he accepts that the main part of the Culm strata and the black Culm limestone are passing conformably into one another. Hence, Murchison infers some part of the black Culm limestone as Coal Measures or Mountain Limestone in age. However, before time step 3, based on his body of evidence, there are no inferences to make regarding the Non-Culm strata. Dating at least some Non-Culm strata as Coal Measures, Murchison would maximize evidential support using some other confirmation measure than $F_{DOJ}(h, e)$, compare Fig. D.12 and Fig. D.13. However, he actually dates the whole of Non-Culm strata as Cambrian in age. There are at least two reasons for this belief change not maximizing evidential support. First, at that time, dating the Non-Culm strata as Cambrian is uncontroversial. Second, his actual dating hypothesis, stating a major gap in the geological sequence of Devon, challenges De la Beche's competence in the field. Hence, for Murchison, it delivers a good opportunity to distinguish himself. At time step 5, dating the whole Non-Culm strata as Old Red Sandstone, Murchison would maximize evidential support using $F_{DOJ}(h, e)$, see Fig. D.14. Note that the final consensus dating hypothesis is part of this set. However, he shows some conservatism in still dating some Non-Culm

[13] Differences between confirmations measures in maximizing evidential support refer to differences in their definitions, compare sec. D.4

[14] For all time steps and persons with a dating hypothesis not maximizing evidential support, the sets of dating hypotheses maximizing evidential support are shown in sec. D.5.

strata as Cambrian.

Entering the stage, Lyell shares the belief that the main part of the Culm strata and its black limestone are passing conformably into one another. Further, he believes in his own principle, which - together with some auxiliary assumptions - places a lower limit on the age of the main part of the Culm strata, namely being older than Old Red Sandstone. Until time step 6, based on his body of evidence, there are no inferences to make regarding the Non-Culm strata. This whole time span, there is a variety of ways to maximize evidential support using some other confirmation measure than $F_{DOJ}(h, e)$. For the first two time steps, presuming the youngest part of the Non-Culm strata as old as the black Culm limestone, Lyell would maximize evidential support, compare Fig. D.11 and Fig. D.10. For the next three steps, Lyell would maximize evidential support by presuming (i) the Non-Culm strata only to be Cambrian in age and (ii) the geological sequence of Devon to be unbroken. Using $Z_{DOJ}(h, e)$, an additional condition applies, namely, dating at least some part of the main part of the Culm as Coal Measures, compare once more Fig. D.11 and Fig. D.10. At time step 3, using $F_{DOJ}(h, e)$, evidential support could be maximized. It shows that, at this time step, the same set of hypotheses maximizes evidential support for both relevance confirmation measures, compare Fig. D.21. However, Lyell renounces to maximize evidential support by following Murchison's dating.

In the beginning, Phillips fails to maximize evidential support using the absolute confirmation measure. At time step 2, he only restricts the space of dating hypotheses by the shared belief that the main part of the Culm passes conformably into the older black Culm limestone. Dating the whole Culm as Old Red Sandstone or even younger, Phillips would maximize evidential support using the absolute confirmation measure, compare Fig. D.17. Using $Z_{DOJ}(h, e)$, there is a totally different set of hypotheses maximizing evidential support, compare Fig. D.18. All of these hypotheses (i) date the Culm strata older than Old Red Sandstone and (ii) deny all Devonian strata being of the same age. However, Phillips renounces to maximize evidential support by following De la Beche in dating all Devonian strata as Cambrian. At the next time step, Phillips additionally presumes a certain temporal order, namely the main part of the Culm being the youngest and the Non-Culm strata being the oldest Devonian strata. Phillips would maximize evidential support by presuming (i) the Non-Culm strata only to be Cambrian in age and (ii) the geological sequence of Devon to be unbroken. Using $Z_{DOJ}(h, e)$, there is one additional condition, namely dating at least some part of the main part of the Culm strata as Coal Measures,

compare once more Fig. D.17 and Fig. D.18. At time step 3, using $F_{DOJ}(h, e)$, evidential support could be maximized. It shows that, at this time step, the same set of hypotheses maximizes evidential support using a relevance confirmation measure, compare Fig. D.21. However, Phillips renounces to maximize evidential support by following Murchison in dating the Culm as Coal Measures and the Non-Culm as Silurian as well as Cambrian in age. At time step 4*b*, Phillips accepts a certain fossil criterion and changes his dating hypothesis, thereby succeeding in maximizing evidential support.

From the beginning until time step 6, Sedgwick fails to maximize evidential support using some other confirmation measure than $F_{DOJ}(h, e)$. During this whole time span, sets of dating hypotheses maximizing evidential support stay the same, compare Fig. D.16 and Fig. D.15. Based on his body of evidence, there are only very few inferences to make regarding the age of some strata. Rejecting all mineralogical and fossil criteria as well as Lyell's principle, Sedgwick only restricts the space of dating hypotheses by presuming a certain temporal order, namely the main part of the Culm being the youngest and the Non-Culm being the oldest Devonian strata. Dating the whole Non-Culm strata as Cambrian and presuming a conformable sequence, Sedgwick would maximize evidential support using the absolute confirmation measure, see Fig. D.15. Using $Z_{DOJ}(h, e)$, there applies one additional condition, namely, dating at least some part of the main part of the Culm strata as Coal Measures in age, see Fig. D.16. At time step 3, using $F_{DOJ}(h, e)$, evidential support could be maximized. It shows that, at this time step, the same set of hypotheses maximizes evidential support using a relevance confirmation measure, compare Fig. D.21. However, Sedgwick renounces to maximize evidential support by following Murchison in dating some of the Non-Culm strata as Silurian in age as well as stating a gap in the geological sequence of Devon.

From time step 3 until 6, De la Beche also fails to maximize evidential support in terms of relevance confirmation. Except for time step 6, sets of dating hypotheses maximizing evidential support are the same for both relevance confirmation measures, see sec. D.5. At time step 6, these two sets differ in dating some of the Non-Culm strata as Cambrian, compare Fig. D.9 and Fig. D.8. They agree upon there being no Non-Culm strata which are Silurian in age. However, De la Beche refuses to maximize evidential support by holding on to this dating. At time step 3, only dating *some* Culm strata as old as the Non-Culm strata and dating some Culm strata as Coal Measures, would maximize evidential support. At time step 5, in or-

der to maximize evidential support, De la Beche has to change his attitude towards there being (*i*) some Non-Culm strata which are Silurian in age and (*ii*) some Culm strata which are Coal Measures in age. However, De la Beche refuses until the final time step a Coal Measures dating of some Culm strata. There are at least two reasons for this belief change not maximizing evidential support. First, before the great Devonian controversy, dating Culm strata older than Old Red Sandstone is uncontroversial. Therefore, conservatism is one possible reason for his belief change not maximizing evidential support. Second, dating some Culm strata as Coal Measures has been first proposed by Murchison using characteristic fossils. Rejecting dating by means of characteristic fossils, De la Beche has no reason to believe in a Coal Measures dating of some Culm strata. Finally, Murchison delivered its proposal as an attack on De la Beche's competence in the field. To secure his livelihood, De la Beche has to refute this attack.

Summing up, most of the time, individual belief changes are rational according to one of the two previously introduced epistemic rules of forming beliefs according to confirmation. However, there are also individual belief changes which are not rational. Presuppose that participants of the great Devonian controversy are rational. Together, it follows that there are exceptions to these epistemic rules. This confirms Feyerabend (1976), stating that there is no scientific rule without any exceptions. He argues for this thesis, for example, referring to the history of science, claiming that for every epistemic rule, there is a historic case where it has been violated, compare (Feyerabend, 1976, p. 35).[15]

Example 3.3.2: Argument with History

(FH1) For all possible epistemic rules: The epistemic rule applies to historic cases.
(FH2) For all possible epistemic rules: There is some historic case where the epistemic rule has been violated.
—
(EM) For all possible epistemic rules: There is some situation where the epistemic rule is violated.

$FH1$ is a shared belief of some important philosophers, see for example Lakatos (1970) and Kuhn (1983). However, there are also others, see for example Carnap

[15]Here, I take it that, given its context, "Grundsatz für das Betreiben von Wissenschaft" can be translated and interpreted as "epistemic rule".

(1950). The recent results relate to $FH2$ by providing conforming instances: For the great Devonian controversy, there are times where participants are not rational, at least in the sense of one of the two previously introduced epistemic rules. Despite providing conforming instances of $FH2$, this thesis must not support relativism in the sense of "anything goes", compare sec. 2.2.2.

Why should one suppose that the main participants of the great Devonian controversy are rational in their individual belief changes, at least most of the time? There is at least one good reason, namely their approachment of a final consensus with total and remarkably high similarity of dating hypotheses and bodies of evidence, respectively.[16] The next chapter analyses approachment to the final consensus for every main participant. Of special interest is the consensus-conduciveness of increase respectively maximization of evidential support. Hence, for the great Devonian controversy, it is analyzed, if a participant's not being rational in the sense of one of the two previously introduced epistemic rules fosters his approaching the final consensus.[17]

[16] Certainly, there is more about scientific rationality than approaching a final consensus. Take as an example approaching the truth and take a look in (Betz, 2013).

[17] This relates to $FS2$ as introduced in sec. 2.2.2 by possibly providing conforming instances. In this case, scientific progress corresponds to approachment to the final consensus and a situation is characterized by a dialectical structure and a participant's position.

ALL			
	DOJ	Z	F
0b	{}	{}	{}
1a	{}	{}	{}
1b	{}	{}	{}
2a	{}	{}	{}
2b	{(DLB, 99.85)}	{(DLB, 99.88)}	{(DLB, 34.89)}
3a	{(MUR, 32.72)}	{(MUR, 36.56)}	{(MUR, 9.79)}
3b	{}	{}	{}
4a	{}	{}	{}
4b	{}	{}	{(DLB, 3.24)}
5a	{(DLB, 15.72), (MUR, 62.79), (LYE, 38.7), (PHI, 30.48), (SED, 18.85)}	{(DLB, 15.06), (MUR, 62.87), (LYE, 38.51), (PHI, 30.39), (SED, 17.62)}	{}
5b	{(DLB, 2.45), (MUR, 61.95), (PHI, 4.1)}	{(DLB, 2.6), (MUR, 63.21), (PHI, 4.13)}	{(MUR, 6.4)}
6a	{(DLB, 26.24), (MUR, 44.88), (LYE, 60.77), (PHI, 40.), (SED, 27.38)}	{(DLB, 25.6), (MUR, 46.05), (LYE, 60.79), (PHI, 39.88), (SED, 26.27)}	{(MUR, 4.03)}
6b	{}	{}	{}
7a	{(DLB, 71.02), (AUS, 73.17)}	{(DLB, 73.6), (AUS, 73.53)}	{(DLB, 15.9), (AUS, 2.6)}
7b	{}	{}	{}
8a	{}	{}	{}
8b	{}	{}	{}

Figure 3.15.: Belief changes are listed in temporal order which decrease evidential support using $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$, respectively. Here, a decrease of less than one per cent is considered as negligible. For some point (x,y), x and y refer to some person and the relative decrease in evidential support, respectively.

ALL			
	DOJ	Z	F
0b	{}	{}	{}
1a	{}	{}	{}
1b	{(MUR, 600.), (LYE, 600.)}	{(MUR, 601.03), (LYE, 601.03)}	{}
2a	{(MUR, 600.), (LYE, 600.)}	{(MUR, 598.48), (LYE, 504.11)}	{}
2b	{(MUR, 600.), (LYE, 600.), (PHI, 75.)}	{(MUR, 598.48), (LYE, 504.11), (PHI, 50.02)}	{}
3a	{(MUR, 600.), (LYE, 600.), (PHI, 75.)}	{(MUR, 597.14), (LYE, 523.15), (PHI, 50.02)}	{}
3b	{(LYE, 100.), (PHI, 100.), (SED, 100.)}	{(DLB, 1.22), (LYE, 101.17), (PHI, 103.22), (SED, 101.22)}	{(DLB, 2.37), (LYE, 1.13), (PHI, 3.09), (SED, 1.17)}
4a	{(LYE, 100.), (PHI, 100.), (SED, 100.)}	{(DLB, 1.3), (LYE, 100.38), (PHI, 103.42), (SED, 100.39)}	{(DLB, 2.51), (PHI, 3.27)}
4b	{(LYE, 100.), (SED, 100.)}	{(DLB, 3.6), (LYE, 100.32), (SED, 100.39)}	{(DLB, 6.98)}
5a	{(DLB, 89.74), (MUR, 300.), (LYE, 300.), (PHI, 89.74), (SED, 100.)}	{(DLB, 95.55), (MUR, 300.93), (LYE, 305.48), (PHI, 89.41), (SED, 100.1)}	{(DLB, 7.49), (LYE, 2.68)}
5b	{(LYE, 100.), (SED, 100.)}	{(DLB, 3.93), (MUR, 2.1), (LYE, 100.09), (SED, 100.1)}	{(DLB, 7.69), (MUR, 4.12)}
6a	{(LYE, 600.), (SED, 100.)}	{(DLB, 4.08), (MUR, 4.77), (LYE, 604.49), (SED, 100.05)}	{(DLB, 8.07), (MUR, 9.41), (LYE, 2.77)}
6b	{(SED, 100.)}	{(SED, 100.05)}	{(DLB, 3.51)}
7a	{(SED, 100.)}	{(SED, 100.05)}	{(DLB, 13.13), (AUS, 1.81)}
7b	{}	{}	{}
8a	{}	{}	{}
8b	{}	{}	{}

Figure 3.16.: Belief changes are listed which do not maximize evidential support using $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$, respectively. Here, a relative difference of less than one per cent is considered as negligible. For some point (x,y), x and y refer to some person and the relative difference in evidential support, respectively.

3.4. Roads to the Final Consensus

This chapter analyses approachment to the final consensus. Of special interest is the consensus-conduciveness of belief changes increasing or maximizing evidential support. The main results of these analyses are shortly summarized in the following.

Roads to the final consensus differ remarkably. There are no two persons following the same road. However, there are rather strong similarities. For all persons, it holds:

- ($H5.1$) Approaching alternates with distancing the final consensus. This is in line with Betz (2013) investigating consensus-conduciveness of controversial debates by means of multi-agent simulations.
- ($H5.2$) Approaching the final consensus in terms of dating hypotheses does not imply an approachment in terms of bodies of evidence, and vice versa. This is a refinement of the analysis of consensus formation given in (Rudwick, 1988).

My analyses show that similarity with the final consensus increases most of the time, but not always. There are two persons never decreasing similarity with the final dating hypothesis, namely Phillips and Austen.

As a result of my analyses, it does not hold that a belief change decreases similarity with the final consensus, iff it decreases evidential support. This is true for all three confirmation measures. There are belief changes decreasing evidential support and not decreasing similarity with the final consensus as well as belief changes decreasing similarity with the final consensus and not decreasing evidential support. However, it shows that, after a sufficiently large number of successive belief changes decreasing evidential support, there is often a considerable change in similarity with the final consensus. Hence, successive decrease of evidential support seems to be a reason for changing beliefs. Note that there are reasons for not changing beliefs, even if evidential support decreases.

As a result of my analyses, it does not hold that a belief change decreases similarity with the final consensus, iff it does not maximize evidential support. This is true for all three confirmation measures. There are belief changes neither maximizing evidential support nor decreasing similarity with the final consensus as well as belief changes decreasing similarity with the final consensus and maximizing evidential

support. Further, it does not hold that a belief change increases similarity with the final consensus, if it maximizes evidential support.

Is maximizing evidential support well designed to approach the final consensus, that is, do most dating hypotheses maximizing evidential support have a sufficiently high degree of similarity with the final consensus? For every time step and person, the ratio of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize evidential is determined. There are several things about this ratio which should be noted:

- ($H5.3$) There are big differences between persons. Hence, there is a dependence between this ratio and a person's body of evidence. However, this dependence cannot be assessed in terms of a bodies similarity with the final consensus. A certain degree of similarity is neither a sufficient nor a necessary condition for the ratio being greater 0.5. Hence, some evidential claims seem to have more impact than others.

- ($H5.4$) There are differences between confirmation measures. In order to maximize the ratio, $Z_{DOJ}(h, e)$ is better than the other two confirmation measures.

- ($H5.5$) For all three confirmation measures and persons, before time step $4a$, it holds: The ratio is less or equal 0.5. Hence, during early phases of the debate, maximizing evidential support is not well designed to approach the final consensus, independent of a certain person.

- ($H5.6$) For all three confirmation measures and most main participants, from time step $7b$ till the end, the ratio equals 1. At the final step, this is true for all persons. Hence, the debate is closed when, as a result of argumentation, evidence accumulation and belief changes, maximizing evidential support is maximally well designed in approaching the final dating hypothesis, independent of a certain person.

The two latter results illustrate the compromise model as introduced in (Kitcher, 1993) in a new way. Presume that participants of the great Devonian controversy undergo the process of maximizing evidential support. As my analyses show this is true at least most of the time, compare Tab. 3.3 and Fig. 3.16. Presume further that cognitive progress is considered as approaching the final consensus. Then, the two latter results corresponds to commitments $C4$ and $C5$ of the compromise model.

3.4.1. Similarity with the Final Consensus

Let us briefly recapitulate what the final consensus is all about, compare Fig. 3.7. All main participants of the great Devonian controversy agree on a certain dating hypothesis, namely the main part of the Culm, its black limestone and the Non-Culm being Coal Measures, Mountain Limestone and Old Red Sandstone in age, respectively. There is no total agreement on a certain body of evidence. However, at the final time step, the intersection of all bodies of evidence is quite large.[18]

Here, approachment to the final consensus dating hypothesis is assessed in terms of the similarity between a participant's dating hypothesis at a certain time step and the final dating hypothesis. Regarding approachment to the final consensus body of evidence, things are a little bit more complicated. First, in my reconstruction, a body of evidence is as small as possible, that is, it does not include sentences which are implied by other evidential beliefs and the dialectical structure. Second, not all sentences of the final consensus are part of the dialectical structure at each time step. Therefore, I take it that the approachment to the final consensus body of evidence is only adequately assessed by the similarity between the deductive closure of a person's body of evidence at a certain time step and the deductive closure of the final consensus body of evidence restricted to those sentences which are part of the dialectical structure at the certain time step. In both cases, that is approachment to the final consensus dating hypothesis and approachment to the final consensus body of evidence, the used similarity measure has already been introduced in sec. 3.2.2.

Definition 3.4.1: Similartiy with the final consensus (CON)

For some person j and time step t, similartiy with the final consensus is assessed in terms of CON_h and CON_e with

$$CON_h = SIM(h_j^t, h_{FIN})$$
$$CON_e = SIM(\overline{e_j^t}, \overline{e_{FIN}} \cap s^t)$$

Here, $\overline{x}$ refers to the deductive closure of x and and s^t refers to the set of all sentences contained in the dialectical structure at time t. Further, h_{FIN} and e_{FIN} refer to the dating hypothesis and the evidential beliefs finally shared by all main participants, respectively.

[18] For example, there is a total agreement on a certain fossil criterion as well as a certain temporal order of all the older strata in Devon. For more examples see Fig. E.1.

Roads to the final consensus differ remarkably. There are no two persons following the same road, see Fig. 3.17. However, there are rather strong similarities. For all persons, it holds:

- ($H5.1$) Approaching alternates with distancing the final consensus. This is in line with Betz (2013) investigating consensus-conduciveness of controversial debates by means of multi-agent simulations.[19]
- ($H5.2$) Approaching the final consensus in terms of dating hypotheses does not imply an approachment in terms of bodies of evidence, and vice versa. This is a refinement of the analysis of consensus formation given in (Rudwick, 1988).

Some general remarks on similarity with the final consensus. For every person and time step, exact values of the similarity with the final dating hypothesis and body of evidence are listed in Fig. E.2 and Fig. E.3, respectively. Similarity with the final dating hypothesis ranges from 0.53 to 1. Similarity with the final consensus body of evidence ranges from 0.42 to 0.97 and offers a much broader spectrum than similarity with the final dating hypothesis.[20] The dating hypothesis most unsimilar to the final dating hypothesis belongs to De la Beche from time step $4b$ until $6a$. In fact, until the final step, no one has a lower degree of similarity with the final dating hypothesis than De la Beche. There is only one person which joins De la Beche, namely Phillips, but only until time step $4a$. Murchison and Lyell accept the final consensus dating hypothesis one step before the final consensus. Regarding similarity with the final dating hypothesis, the spectrum is widest for De la Beche and narrowest for Austen.[21] The same is true regarding the similarity with the final consensus body of evidence.[22] For Phillips, the spectra of similarity with the final dating hypothesis is remarkably wide and the spectra of similarity with the final consensus body of evidence is remarkably narrow. The body of evidence most unsimilar to the final consensus belongs to De la Beche entering the stage. In fact, the 4 bodies of evidence which are most unsimilar to the final consensus belong to De la Beche from time step $0b$ until $2a$. However, at time step $2b$, with his drastic

[19] Further, my result rules out such models of rational consensus formation as for example presented in (Hegselmann and Krause, 2002).

[20] In the first case there are 70 different values, in the second case, there are only 7.

[21] Width of spectra corresponding to De la Beche, Phillips, Sedgwick, Lyell, Murchison and Austen: $0.47, 0.40, 0.33, 0.33, 0.33, 0.13$.

[22] Width of spectra corresponding to De la Beche, Lyell, Murchison, Sedgwick, Phillips and Austen: $0.54, 0.30, 0.27, 0.18, 0.12, 0.08$.

change of individual beliefs, De la Beche joins Lyell and Murchison in similarity with the final consensus body of evidence.

Some more general remarks on the dynamics of similarity with the final consensus. Remember that dating hypotheses are changed individually. Therefore, similarity with the final dating hypothesis only changes at time steps t_b. Remember also that, for every time step t_a except $6a$, there are new shared background beliefs. A new shared background belief is an additional agreement with the final consensus causing an increase in similarity with the final consensus body of evidence. For every time step t_a, there are some new sentences and arguments. Most of the time, the dialectical structure and a person's body of evidence do not imply a new sentence or its negation. Hence, most of the time, a person suspends judgment on a new sentence causing a decrease in similarity with the final consensus body of evidence.

Are there belief changes such that similarity with the final consensus decreases? My analyses show that, for some time steps and persons, similarity with the final consensus decreases, see Fig. 3.17 and Fig. 3.18. However, most of the time, similarity with the final consensus increases. This is true for similarity with the final dating hypothesis as well as for similarity with the final consensus body of evidence, compare Tab. 3.4. There are two persons never decreasing similarity with the final dating hypothesis, namely Phillips and Austen. The average percentage decrease of similarity with the final dating hypothesis is much greater than that of similarity with the final consensus body of evidence, namely 11.5 and 6.0.

In the following, for every person, I shortly delineate the road to the final consensus. De la Beche starts with a rather low degree of similarity with the final consensus. There is only one belief change decreasing his dating hypothesis' similarity with the final dating hypothesis, namely at time step $4b$. Here, De la Beche adopts Murchison's dating of some Non-Culm strata as Silurian, see for example Fig. B.3. At time step $6b$, De la Beche completely compensates this loss by dating some Non-Culm strata as Old Red Sandstone. There are several belief changes decreasing the similarity between his body of evidence and the final consensus body of evidence. At time step $1a$, with Murchison and Lyell proposing a new temporal order of all the older strata of Devon, namely the main part of the Culm strata being the youngest, the similarity with the final consensus body of evidence decreases drastically. However, for the following five belief changes, it increases. Increase is especially large for De la Beche's individual belief changes at time steps 2 and 3. At time step $2b$, De

la Beche adopts Phillips's objection against all previously discussed mineralogical and fossil criteria, namely that there are always local variations in sedimentation as well as in fauna and flora. At time step 3*b*, as a result of Murchison's Devon campaign, De la Beche changes his mind about the geological structure of Devon. He now joins Murchison in considering the main part of the Culm as the youngest Devonian strata. For the following two shared belief changes, namely 4*a* and 5*a*, the similarity between De la Beche's body of evidence and the final consensus body of evidence slightly decreases. At these time steps, there are no additional contradictions but percentage increase is considerably larger for judgment suspensions than agreements, compare Fig. E.4. At time step 6*b*, changing his mind about the temporal order of all the older strata in Devon once again, De la Beche causes a drastic decrease in similarity with the final consensus body of evidence. Here, De la Beche considers some south Devonian Non-Culm strata as Old Red Sandstone in age. Together with dating the Culm as Silurian in age, this implies a new temporal order, namely the main part of the Culm strata being older than some south Devonian Non-Culm strata. De la Beche's similarity with the final consensus body of evidence decreases further by not classifying fauna and flora of the Non-Culm strata as a local variation. For the following two belief changes, namely 7*a* and 7*b*, it slightly decreases. At time step 7*a*, De la Beche contradicts the final consensus by rejecting that most fossils of the main part of the Culm are known from Coal Measures strata as well as by accepting some new empirical evidence confirming his dating hypothesis. At time step 7*b*, De la Beche's similarity with the final consensus body of evidence slightly decreases due to rejecting some assumptions about the fossil content of the Non-Culm strata.

Murchison starts somewhere in the middle of the similarity spectra. There is only one belief change decreasing his dating hypothesis' similarity with the final dating hypothesis, namely at time step 3*b*. Here, as a result of his Devon campaign, Murchison dates some Non-Culm strata as Silurian, see for example Fig. B.3. At time steps 5*b* and 6*b*, Murchison compensates and overcompensates this loss by dating some Non-Culm strata as Old Red Sandstone and the black Culm limestone as Mountain Limestone in age, respectively. Murchison agrees with the final consensus dating hypothesis already at the penultimate time step. There are several belief changes decreasing the similarity between his body of evidence and that of the final consensus. At time step 4*a*, denying any local variations in fauna and flora, it decreases

considerably.[23] It decreases further for the following two belief changes. At time step 4*b*, rejection of a newly introduced fossil criterion, namely a limited version of the characteristic fossil assemblage principle, leads to a slight decrease in similarity. Time step 5*a* is all about dating some north Devonian Non-Culm strata. Here, Murchison contradicts the final consensus by not considering the Non-Culm strata as a local variation. With the next individual belief change, giving up his belief in dating by means of characteristic fossils, there comes a major increase in similarity with the final consensus body of evidence. At time step 6*b*, similarity decreases once again considerably. Here, Murchison changes his beliefs fundamentally, returning to dating by means of characteristic fossils and rejecting the latest fossil findings in some Non-Culm strata. This way, Murchison contradicts the final consensus 8 times more often than before. As a consequence, Murchison also contradicts an assumption about the fossil content of Non-Culm strata newly introduced at time step 7*a*. With the next individual belief change, there comes a drastic increase in similarity with the final consensus body of evidence. Here, Murchison gets rid of all contradictions with the final consensus, that is 17 contradictions. Further, there is a major percentage decrease in judgment suspensions and increase in agreements, compare Fig. E.4. For example, Murchison changes his mind regarding fossil criteria, rejecting dating by means of characteristic fossils and accepting a limited version of the Lyellian principle. Another example is his change of mind regarding fossil findings in some Non-Culm strata. At time step 8*a*, there is a minor decrease in similarity with the final consensus body of evidence, due to percentage increase being considerably larger for contradictions than agreements, compare Fig. E.4. At this time step, Murchison holds on to belief in the Non-Culm strata being no local variation. Together with his other beliefs, he infers the existence of Old Red Sandstone fish fossils in some Non-Culm strata. As a consequence, Murchison rejects results of Phillips' Devon campaign amassing and statistically classifying fossils.

The dynamics of Lyell's dating hypothesis' similarity with the final dating hypothesis are those of Murchison, except at time steps 5*b* and 6*a*. Here, Lyell does not join Murchison in dating some north Devonian Non-Culm strata as Old Red Sandstone in age. Regarding similarity with the final consensus body of evidence, Lyell's and Murchison's dynamics are quite similar. However, there are some considerable differences, namely at time steps 5*b*, 6*a* and 8*a*. At time steps 5*b* and 6*a*, holding

[23] As a consequence of there being no local variations in fauna and flora, Scottish Old Red Sandstone strata serve as a blueprint for Old Red Sandstone strata elsewhere.

on to his principle, Lyell has to deny (*i*) there being any local variations in fauna and flora and (*ii*) Scottish Old Red Sandstone strata supporting only a few peculiar fish fossils not known from other strata. Further, in order to consistently hold on to his dating hypothesis, Lyell denies fossil findings in north Devonian Non-Culm strata as well as in the black Culm limestone. As a consequence, in these cases, similarity with the final consensus body of evidence is considerably smaller for Lyell than for Murchison. At time step 8*a*, still holding on to his principle, Lyell's body of evidence is no longer dialectically consistent. As a consequence, its similarity with the final consensus body of evidence cannot be determined and does not figure in Fig. 3.17.

As already mentioned, Phillips's similarity dynamics are quite remarkable. First, the spectrum of similarity with the final dating hypothesis is remarkably wide and the spectrum of similarity with the final consensus body of evidence is remarkably narrow. Second, similarity with the final dating hypothesis never decreases. Mimicking De la Beche, Phillips starts with a rather low degree of similarity with the final dating hypothesis. At time step 4*b*, detaching from De la Beche, Phillips changes his dating hypothesis drastically. He now considers the black Culm limestone as Mountain Limestone and some Non-Culm strata as Silurian in age. The increasing effect of the first belief change outweighs the decreasing effect of the second one. Phillips starts with a remarkably high degree of similarity with the final consensus body of evidence, not least because of there being no contradictions with the final consensus. During the whole time span of the debate, there are remarkably few contradictions with the final consensus, compare Fig. E.4. With the first shared belief change, the similarity of Phillips's body of evidence with the final consensus decreases due to a larger percentage increase of judgment suspensions than agreements, compare Fig. E.4. This is also true for shared belief changes at time steps 4*a*, 5*a* and 6*a*. At time step 3*a*, the decrease is so small that Fig. 3.17 shows no difference. At time step 3*b*, adopting Murchison's view of the geological structure of Devon, similarity with the final consensus body of evidence increases considerably. Accepting a limited version of the characteristic fossil assemblage principle and a certain assumption about the fossil content of the black Culm limestone, it increases once more at time step 4*b*. Further, Phillip approaches the final consensus by considering Scottish Old Red Sandstone strata as a local variation. At time step 6*b*, rejecting the new fossil finding in some south Devonian Non-Culm strata, Phillips moves away from the final consensus. At time step 7*a*, Phillips contradicts the final

consensus by rejecting a certain assumption about the fossil assemblage of the Non-Culm strata. Moving to time step 7*a* (7*b*), the decrease (increase) of similarity with the final consensus body of evidence is so small that Fig. 3.17 shows no difference.

First, Sedgwick joins Murchison in dating the Culm as Coal Measures and the Non-Culm strata as Silurian and Cambrian in age. At time step 6*b*, Sedgwick detaches from Murchison by changing his mind about the black Culm limestone. He now dates it as being Mountain Limestone as well as Silurian and Old Red Sandstone in age. This belief change decreases his dating hypothesis' similarity with the final dating hypothesis considerably. Sedgwick starts with a high degree of similarity with the final consensus body of evidence. However, at the same time, there are two persons with a degree at least as high as Sedgwick, namely Phillips and De la Beche. For the next three belief changes, similarity with the final consensus body of evidence decreases considerably. At time steps 4*a*and 5*a*, percentage increase is larger for judgment suspensions than agreements, compare Fig. E.4. At time step 4*b*, there are five new contradictions with the final consensus. Sedgwick rejects dating by means of characteristic fossil species assemblages and considers Scottish Old Red Sandstone strata as blueprint for strata of the same era elsewhere. Belief changes at time steps 5*b* and 6*a* changes similarity with the final consensus body of evidence to such a small extent that Fig. 3.17 shows no difference at all. For the two next belief changes, it decreases. At time step 6*b*, Sedgwick rejects new fossil findings from South Devon. As a consequence, Sedgwick also rejects a certain assumption about the fossil species assemblage of the Non-Culm strata at time step 7*a*. The following two belief changes increase respectively decrease similarity with the final consensus body of evidence to the same extent. At time step 7*b*, Sedgwick joins De la Beche in considering the geological sequence of Devon to be unbroken. Before joining the final consensus, Sedgwick has to change his attitude towards fossil findings in the Non-Culm strata as well as the status of Scottish Old Red Sandstone strata.

Austen enters the stage rather late, namely at time step 6*b*. His similarity dynamics are quite simple. Similarity with the final dating hypothesis is very high, right from the start, and never decreasing. Austen always dates the main part of the Culm, its black limestone and some Non-Culm strata as Coal Measures, Mountain Limestone and Old Red Sandstone, respectively, compare Fig. B.3. Similarity with the final consensus body of evidence is also very high from the start, but decreases with every belief change, except the last one. However, it is never really low. Moving from time step 6*b* to 7*a*, percentage increase is larger for judgment suspensions than

agreements, compare Fig. E.4. At time step 7*b*, similarity with the final consensus body of evidence slightly decreases due to rejecting some assumptions about the fossil content of the Non-Culm strata. As a consequence, Austen rejects Phillips's fossils finding in the Non-Culm strata at time step 8*a*.

Note finally that, at a sufficiently late time step, all persons change their beliefs such that evidential support is maximized. From time step 7*b* until the end, all persons maximize their evidential support. This is true for all three confirmation measures. If not using $F_{DOJ}(h, e)$, from time step 6*b* until the end, all persons maximize their evidential support, the only exception being Sedgwick. Using $F_{DOJ}(h, e)$, during the same time span, all persons maximize their evidential support, the only exceptions being De la Beche and Austen.

Combining Fig. 3.17 with results of the last chapter, the consensus-conduciveness of increase and maximization of evidential support can be assessed. For more on this topic, see the next chapter.

	CON_h	CON_e
Individual belief changes	$\frac{4}{41}$	$\frac{20}{41}$
Shared belief changes	–	$\frac{5}{17}$

Table 3.4.: Proportion of belief changes where similarity with the final consensus decreases. Here, a decrease of less than one per cent is considered as negligible.

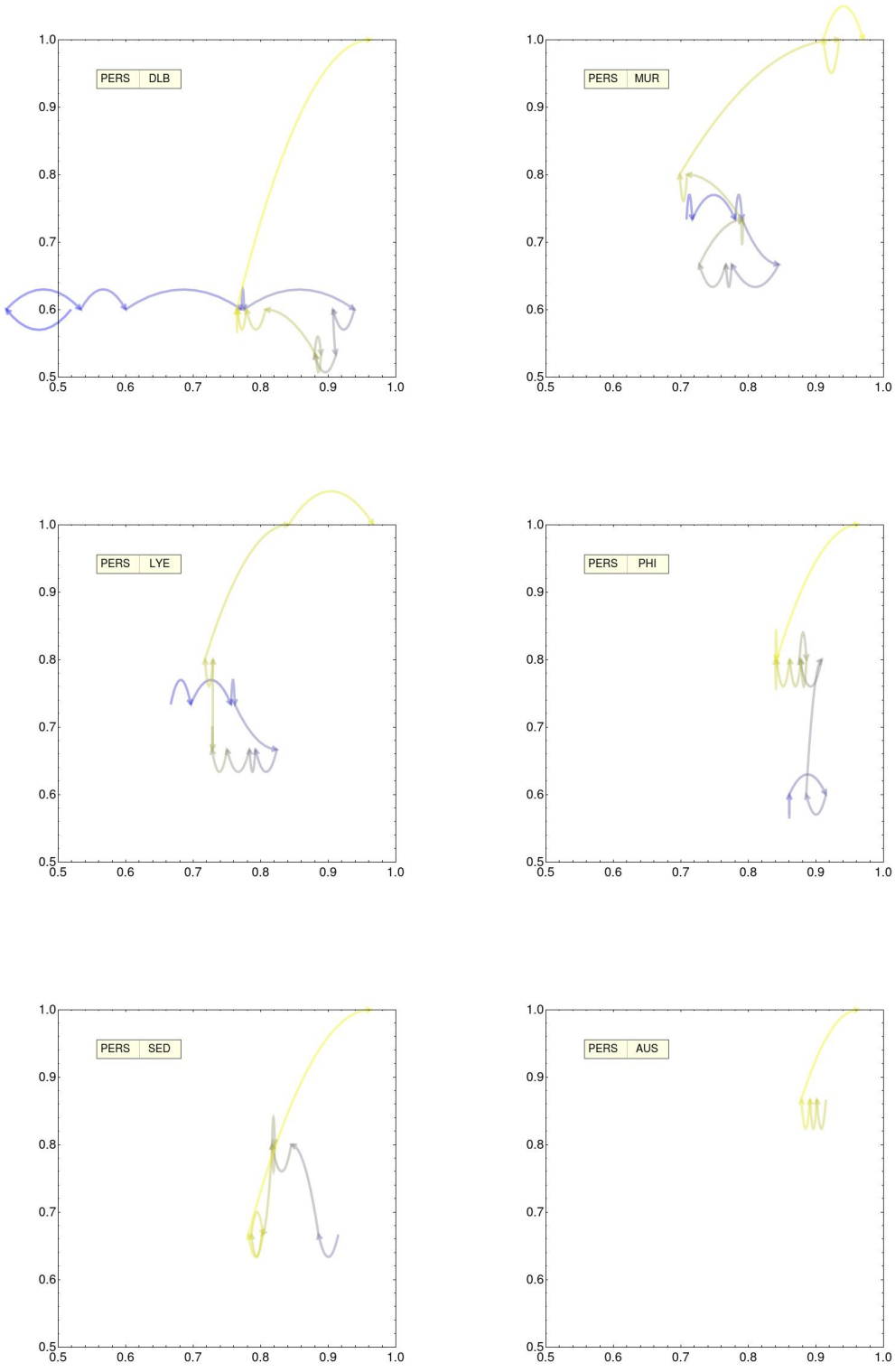

Figure 3.17.: Similarity with the Final Consensus. For every main participant, similarity between her dating hypothesis and the final dating hypothesis is plotted over similarity between her body of evidence and the final consensus body of evidence. Time t is represented using a color function. Turning from blue over green to yellow, t increases.

TIME	DECREASE CON_h	DECREASE CON_e
0b	–	–
1a	–	((DLB, 18.5))
1b	–	–
2a	–	–
2b	–	–
3a	–	–
3b	((MUR, 9.1), (LYE, 9.1))	–
4a	–	((DLB, 3.4), (MUR, 8.3), (LYE, 3.9), (PHI, 3.2), (SED, 3.2))
4b	((DLB, 11.1))	((MUR, 1.1), (LYE, 1.1), (SED, 4.5))
5a	–	((DLB, 3.5), (MUR, 5.1), (LYE, 4.3), (PHI, 3.6), (SED, 3.4))
5b	–	((LYE, 3.))
6a	–	–
6b	((SED, 16.7))	((DLB, 8.4), (MUR, 10.2), (PHI, 1.9), (SED, 1.8))
7a	–	((DLB, 3.5), (MUR, 1.5), (LYE, 1.6), (PHI, 2.3), (SED, 2.), (AUS, 1.5))
7b	–	((DLB, 1.7), (AUS, 1.1))
8a	–	((MUR, 2.5), (SED, 2.7), (AUS, 1.4))
8b	–	–

Figure 3.18.: Belief changes are listed in temporal order which decrease similarity with the final consensus. Here, a decrease of less than one per cent is considered as negligible. For some point (x, y), x refers to some person and y refers to the relative decrease in similarity.

3.4.2. Increasing Evidential Support

Comparing Fig. 3.15 with Fig. 3.17, consensus-conduciveness of increasing evidential support can be assessed. In order to facilitate this comparison, a more illustrative but less exact comparison is given in the following. Fig. 3.19, Fig. 3.20 and Fig. 3.21 show the same as Fig. 3.17, namely, for every main participant, similarity between her dating hypothesis and the final dating hypothesis over similarity between her body of evidence and the final consensus body of evidence. Additionally, those belief changes are marked with a red circle decreasing confirmation. The size of decrease in confirmation, x, relates to the size of a red circle, y. For illustrative reasons, the following shrinking has been performed:

$$y = 0.02 * Log[x] + 0.02 \tag{3.7}$$

As a result of my analyses, it does not hold that a belief change decreases similarity with the final consensus, iff it decreases evidential support. This is true for all three confirmation measures. First, there are belief changes decreasing evidential support and not decreasing similarity with the final consensus. Take as an example De la Beche at time step $2b$. Here, for all three confirmation measures, evidential support decreases drastically. At the same time, the similarity between his body of evidence and the final consensus body of evidence increases considerably. Second, there are belief changes decreasing similarity with the final consensus and not decreasing evidential support. Take as an example De la Beche at time step $1a$. Here, the similarity between his body of evidence and the final consensus body of evidence decreases considerably.[24] At the same time, for all three confirmation measures, evidential support does not change at all. As another example take Murchison at time step $3b$. Here, similarity between his dating hypothesis and the final dating hypothesis decreases considerably. At the same time, for all three confirmation measures, evidential support increases considerably.

Is there some other relation between decreasing evidential support and changing similarity with the final consensus? My analyses show that, after a sufficiently large number of successive belief changes decreasing evidential support, there is often a considerable change in similarity with the final consensus. Take as an example

[24] As for all time steps t_a, there is no change in similarity between a participant's dating hypothesis and the final dating hypothesis. Remember that dating hypotheses are only changed individually, that is at time steps t_b.

Murchison at time steps $3a$ and $3b$. Here, for all three confirmation measures, there is a considerable decrease in evidential support followed by a change of similarity between his dating hypothesis and the final dating hypothesis. Consider also Lyell at time steps $6a$ and $6b$. Here, if not using $F_{DOJ}(h, e)$, there is a considerable decrease in evidential support followed by a change of similarity between his dating hypothesis and the final dating hypothesis. As another example take Sedgwick from time step $5a$ until $6b$. Here, for three successive time steps, namely $5a$, $5b$ and $6a$, evidential support decreases, if not using $F_{DOJ}(h, e)$. At $6b$, there is a considerable change of similarity between his dating hypothesis and the final dating hypothesis. As a last example take De la Beche and the same time span. As for Sedgwick from $5a$ until $6a$, evidential support decreases successively, if not using $F_{DOJ}(h, e)$. At $6b$, there is a considerable change of both types of similarity with the final consensus. Hence, decrease or successive decrease of evidential support seems to be a reason for changing beliefs.

Note that there are reasons for not changing beliefs, even with evidential support decreasing. Take as an example Phillips from time step $5a$ until $6a$. Here, evidential support decreases successively, if not using $F_{DOJ}(h, e)$. However, Phillips does not change beliefs considerably. He holds on to his dating hypothesis, that is similarity between his dating hypothesis and the final dating hypothesis does not change at all. The degree of similarity between his body of evidence and the final consensus body of evidence changes, but only minimally. Phillips accepts that there are some Carboniferous fossils in North Devon, but rejects Austen's assumption about the fossil content of some South Devonian limestones. Before campaigning Devon himself, Phillips is not willing to consider the amounts of collected fossils large enough to safely draw inferences upon. Hence, Phillips's conservatism seems to be based on his scientific skepticism. Consider also Austen and De la Beche at time step $7a$. Here, evidential support decreases. This is true for all three confirmation measures. However, for the next two steps, there is no change in similarity between his dating hypothesis and the final dating hypothesis, and only a minor change in similarity between his body of evidence and the final consensus body of evidence. During the whole debate, De la Beche strives to vindicate his competence in the field having been heavily attacked by Murchison. From time step $6b$ until the final step, De la Beche separates from the others by proposing a new geological structure of Devon and new empirical evidence. So, De la Beche's conservatism seems to be based on his striving for vindication. During the whole debate, Austen strives for a reputation

as a first class geologist, but in vain.[25] So, Austen's conservatism seems to be based on his being on the outside. Note that, for all previously discussed persons and time steps, evidential support is relatively large, compare Fig. 3.12.

There are also other reasons for changing beliefs than stopping decrease in evidential support. Take as an example Lyell at time step 3*b*. Here, Lyell changes his dating hypothesis following Murchison's lead. Another example is Phillips at time step 4*b*. Here, Phillips changes his dating hypothesis as a result of promoting his own scientific views and work.[26] Instead of Cambrian, he now dates the black Culm limestone as Mountain Limestone. As a last example consider De la Beche at time step 2*b*. Here, De la Beche accepts Phillips's objection against all previously discussed mineralogical and fossil criteria, namely that there are always local variations in sedimentation as well as in fauna and flora. This belief change does not stop but causes decrease in evidential support.

Summing up, for all three confirmation measures, it does not hold that a belief change decreases similarity with the final consensus, iff it decreases evidential support. Note that decrease or successive decrease of evidential support is a reason for changing beliefs, that is considerably changing similarity with the final consensus, but only one reason among others. Sometimes, there is a reason for holding on to one's beliefs, although evidential support decreases.

[25] There is at least one geographical reason. At that time, the geological society mostly centers in London. At the same time, Austen spends most of his time at home, that is somewhere in South Devon.

[26] Phillip promotes a differentiated way of dating by means of fossils, namely a limited version of the characteristic fossil assemblage principle. Before the great Devonian controversy, Phillips campaigned Yorkshire and the Pennines, amassing and classifying fossils for example from Mountain Limestone strata.

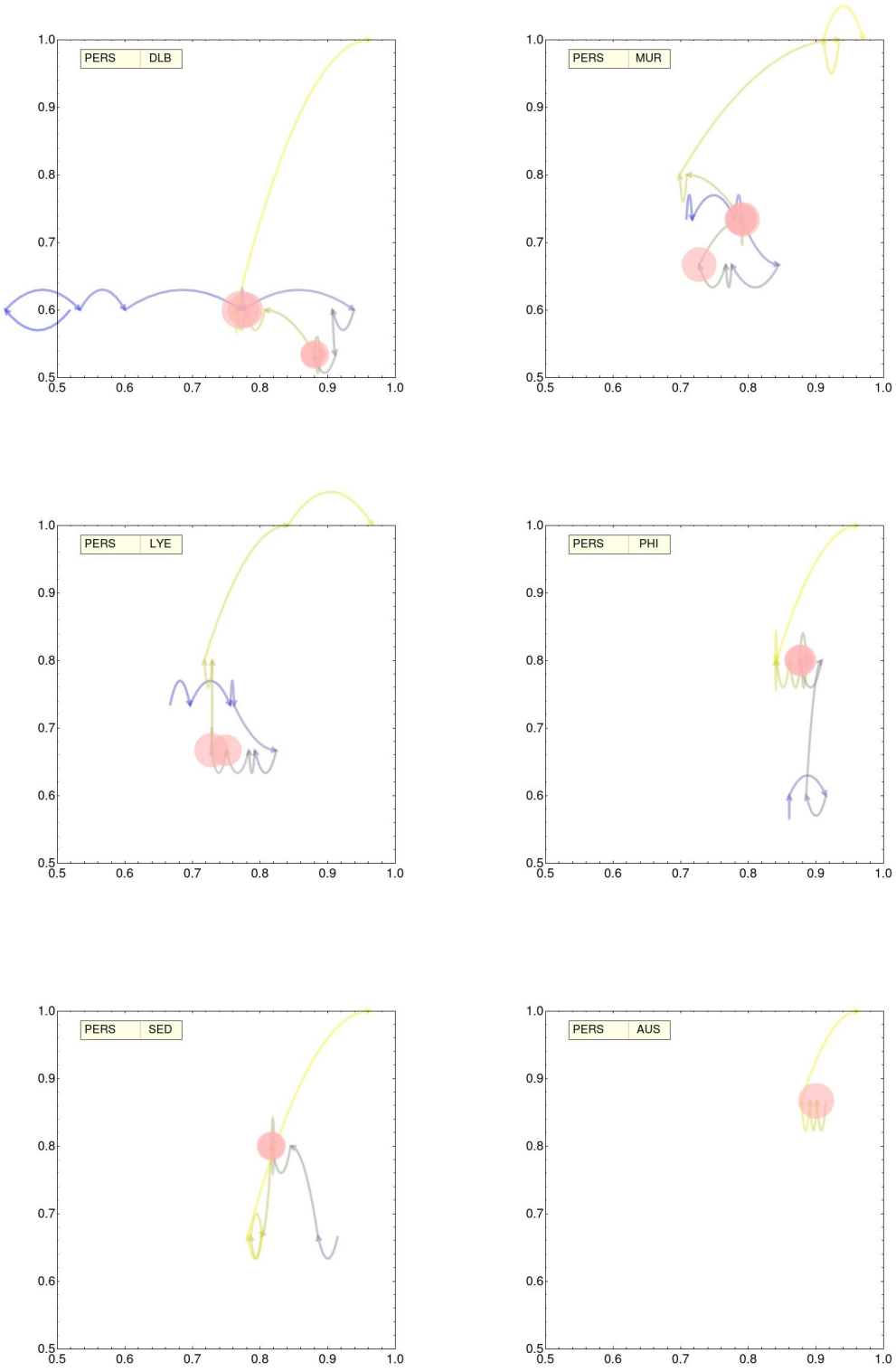

Figure 3.19.: Similarity with the final consensus is shown as in Fig. 3.17. Additionally, those belief changes are marked with a red circle decreasing $DOJ(h|e)$. The size of the red circle relates to the relative decrease of confirmation.

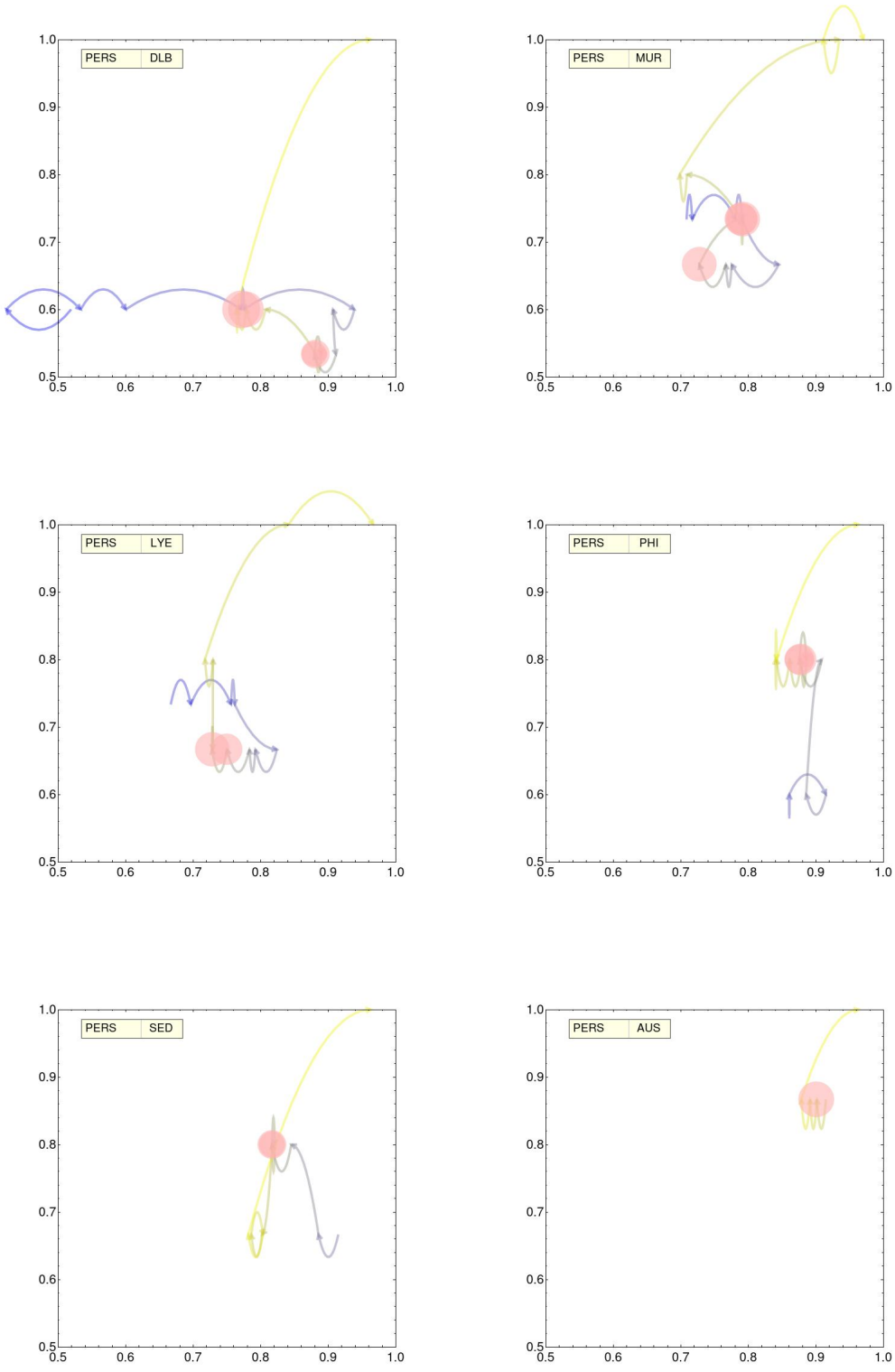

Figure 3.20.: Similarity with the final consensus is shown as in Fig. 3.17. Additionally, those belief changes are marked with a red circle decreasing $Z_{DOJ}(h, e)$. The size of the red circle relates to the relative decrease of confirmation.

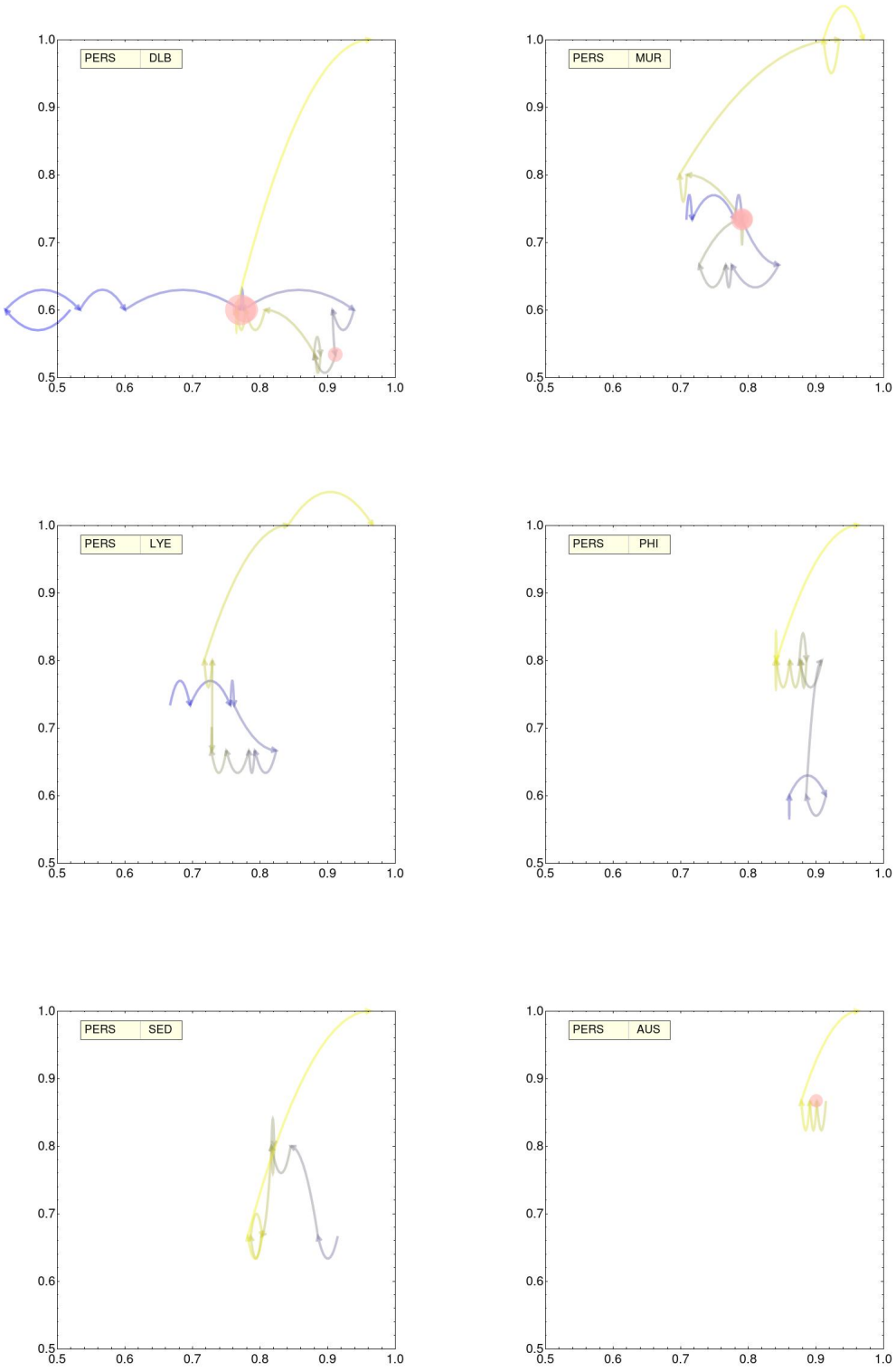

Figure 3.21.: Similarity with the final consensus is shown as in Fig. 3.17. Additionally, those belief changes are marked with a red circle decreasing $F_{DOJ}(h, e)$. The size of the red circle relates to the relative decrease of confirmation.

3.4.3. Maximizing Evidential Support

Comparing Fig. 3.16 with Fig. 3.17, consensus-conduciveness of maximizing evidential support can be assessed. In order to facilitate comparison, a more illustrative but less exact comparison is given in Fig. 3.22, Fig. 3.23 and Fig. 3.24.

Fig. 3.22, Fig. 3.23 and Fig. 3.24 show the same as Fig. 3.17, namely, for every main participant, similarity between her dating hypothesis and the final dating hypothesis over similarity between her body of evidence and the final consensus body of evidence. Additionally, those belief changes are marked with a red circle not maximizing confirmation. The size of the red circle, y, relates to the relative difference between the highest possible degree of confirmation and the actual one, x. For illustrative reasons, the following shrinking has been performed:

$$y = 0.02 * Log[x] + 0.02 \tag{3.8}$$

As a result of my analyses, it does not hold that a belief change decreases similarity with the final consensus, iff it does not maximize evidential support. This is true for all three confirmation measures. First, there are belief changes neither maximizing evidential support nor decreasing similarity with the final consensus. Take as an example Murchison at time step $2b$. Here, not using $F_{DOJ}(h, e)$, evidential support is far from being maximal, but similarity between his body of evidence and the final consensus body of evidence increases. Another example is De la Beche at time step $3b$. Here, using $F_{DOJ}(h, e)$, evidential support is not maximal, but similarity between his body of evidence and the final consensus body of evidence increases considerably. Second, there are belief changes decreasing similarity with the final consensus and maximizing evidential support. Take as an example Austen at time step $8a$. Here, both types of similarity with the final consensus decreases slightly while evidential support is maximal.

As a further result of my analyses, it does not hold that a belief change increases similarity with the final consensus, if it maximizes evidential support. Take as an example Murchison at time steps $6b$, $7a$ and $8a$. Here, evidential support is maximal for all three confirmation measures, but similarity between his body of evidence and the final consensus body of evidence decreases.

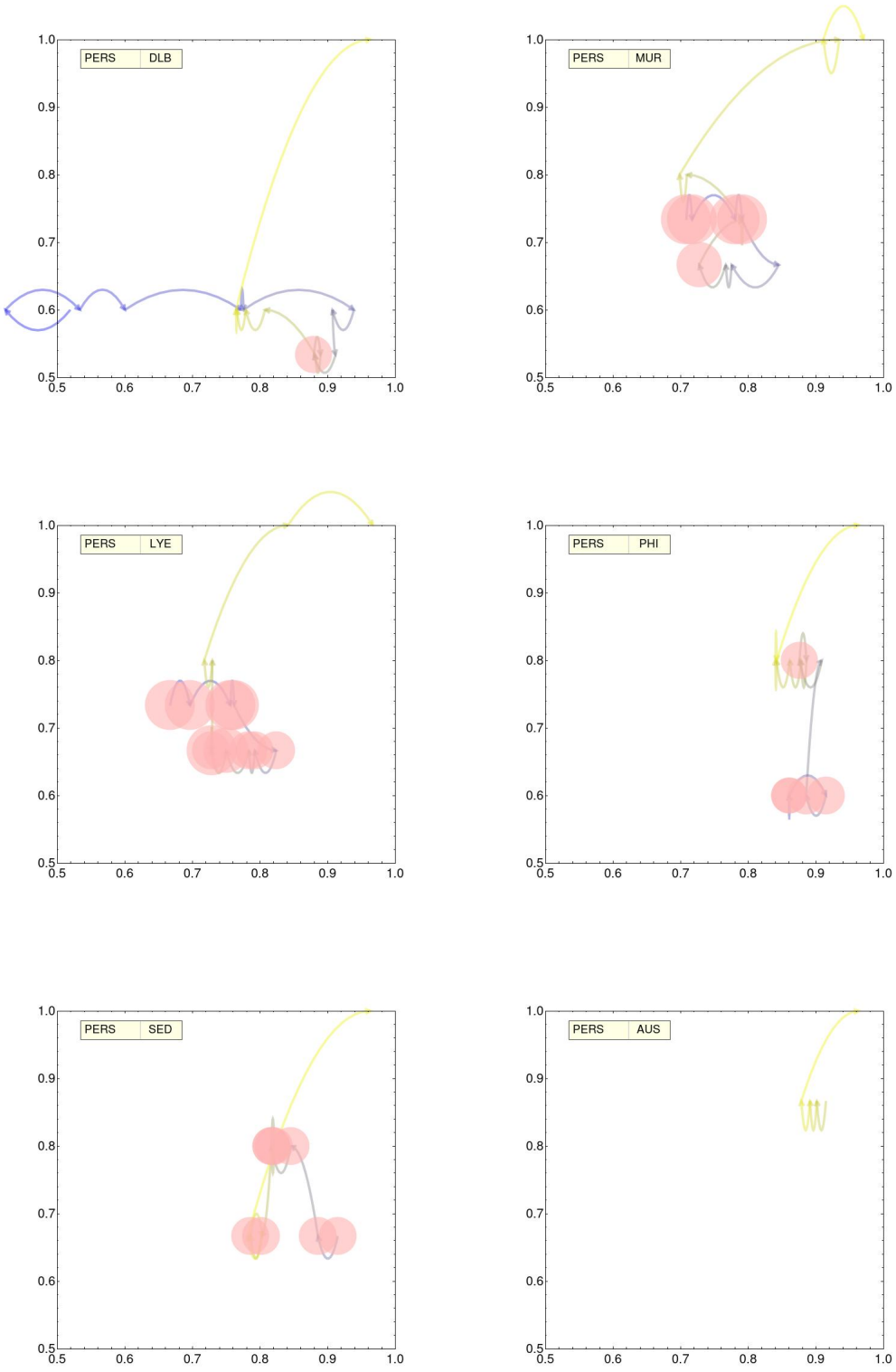

Figure 3.22.: Similarity with the final consensus is shown as in Fig. 3.17. Additionally, those belief changes are marked with a red circle not maximizing $DOJ(h|e)$. The size of the red circle relates to the relative difference between the highest possible degree of confirmation and the actual one.

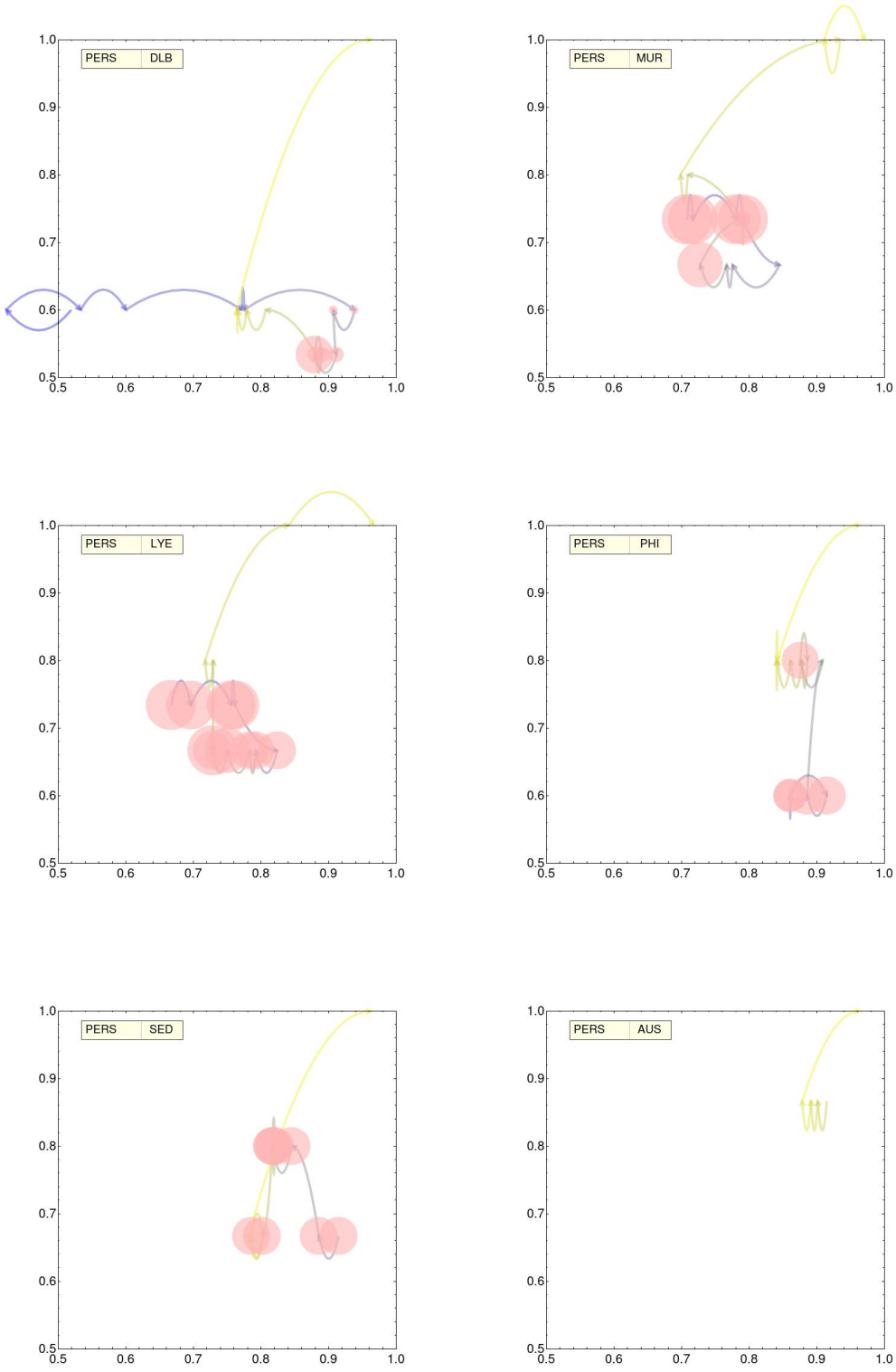

Figure 3.23.: Similarity with the final consensus is shown as in Fig. 3.17. Additionally, those belief changes are marked with a red circle not maximizing $Z_{DOJ}(h, e)$. The size of the red circle relates to the relative difference between the highest possible degree of confirmation and the actual one.

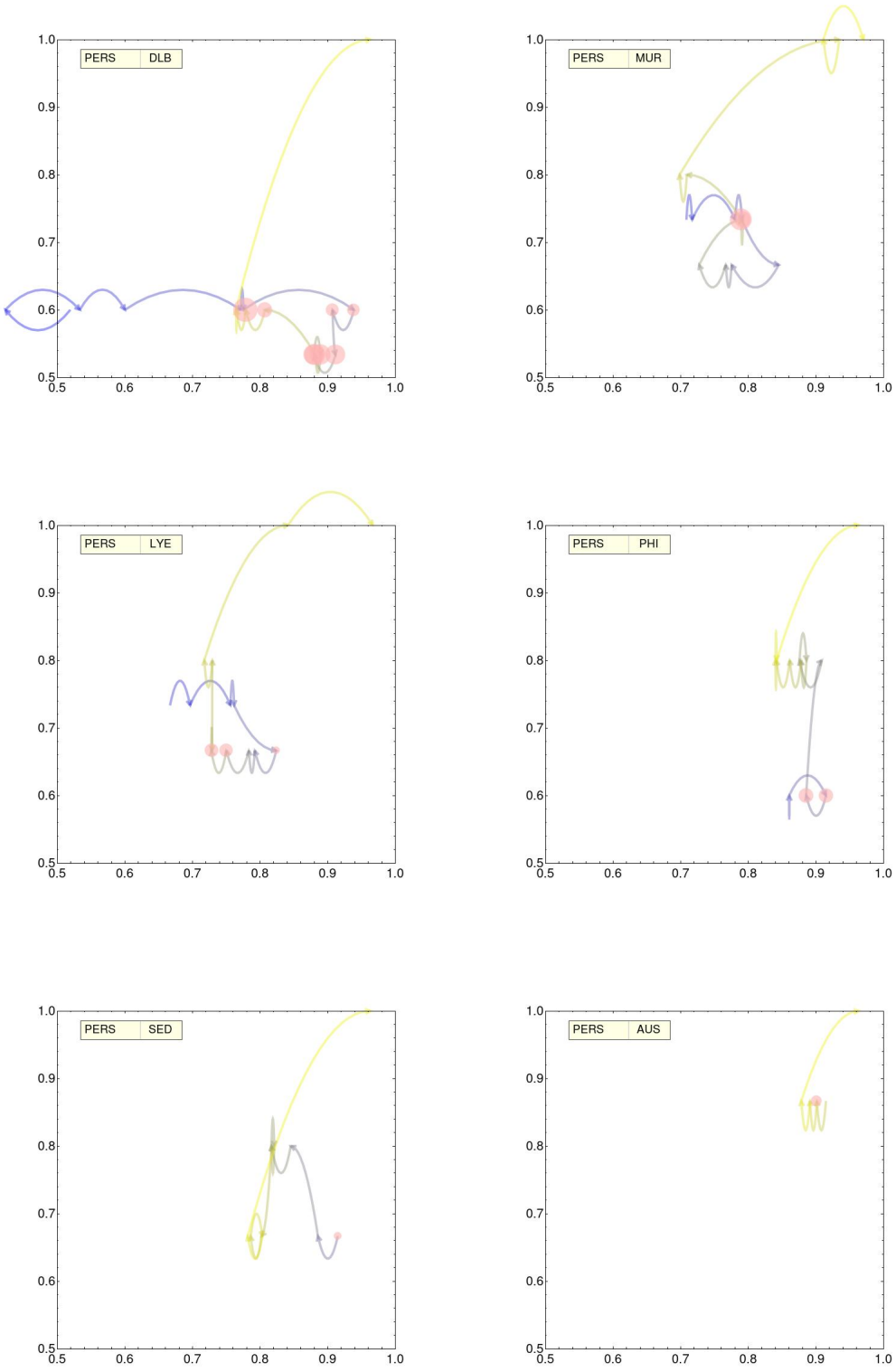

Figure 3.24.: Similarity with the final consensus is shown as in Fig. 3.17. Additionally, those belief changes are marked with a red circle not maximizing $F_{DOJ}(h, e)$. The size of the red circle relates to the relative difference between the highest possible degree of confirmation and the actual one.

Is there some other relation between maximizing evidential support and increasing similarity with the final consensus? For example, do most dating hypotheses maximizing evidential support have a sufficiently high degree of similarity with the final dating hypothesis? Presume that success means accepting a dating hypothesis which is sufficiently similar to the final dating hypothesis.[27] Then, the last question translates into the following one: Is the success ratio of the previously introduced epistemic rule $RAT2$ greater 0.5? The success ratio is given by the relative frequency of dating hypotheses which are sufficiently similar with the final dating hypothesis among all those dating hypotheses maximizing evidential support, and determined via the following steps:

1. For every time step and person, all dating hypotheses maximizing evidential support are determined.

2. For every time step, person and dating hypothesis maximizing evidential support, its degree of similarity with the final dating hypothesis is calculated. Results can be found in sec. E.2.1.

3. For every time step and person, the relative frequency of dating hypotheses which are sufficiently similar with the final dating hypothesis among all those dating hypotheses maximizing evidential support is determined. This is done using results from sec. E.2.1 as well as sec. E.2.2.

Fig. 3.25 shows those cases where the success ratio is greater 0.5. There are several things about this ratio which should be noted.

First, there are big differences between persons. Hence, there is a dependence on the body of evidence. Consider for example the two most extreme cases, namely Sedgwick and Austen. For Sedgwick, the success ratio always equals or is less than 0.5. For Austen the same ratio is always 1. How do bodies of evidence differ? Is there a dependence between the similarity with the final consensus and the success ratio being greater 0.5? Fig. 3.26, Fig. 3.27 and Fig. 3.28 show that a certain degree of similarity between a person's body of evidence and the final consensus body of evidence is neither a sufficient nor a necessary condition for the success ratio being greater 0.5. These plots show the same as Fig. 3.17, namely, for every main participant, similarity between her dating hypothesis and the final dating hypothesis over

[27] Here, I take it that a similarity of 0.8 is sufficient. If some dating hypothesis is similar to the final dating hypothesis with a degree of 0.8, then it supports 3 atomic dating hypotheses contradicting the final consensus.

similarity between her body of evidence and the final consensus body of evidence. Additionally, those time steps are marked with a red circle where the success ratio is greater than 0.5. The actual value of the ratio, x, relates to the size of a red circle, y. For illustrative reasons, the following shrinking has been performed:

$$y = 0.07 * Log[x] + 0.07 \tag{3.9}$$

Take as an example Phillips and absolute confirmation. From time step $4b$ until the penultimate step, the success ratio is greater than 0.5. This is not true from the beginning until time step $4a$. However, during both time spans, similarity between his body of evidence and the final consensus body of evidence is of the same order. As another example take De la Beche and Murchison at time steps $3b$ and $7b$. In both cases, similarity between his body of evidence and the final consensus body of evidence is approximately the same, compare Fig. E.3. But only for Murchison at $7b$, it holds that the success ratio is greater 0.5. This is true for all three confirmation measures. Summing up, it seems that the success ratio's dependence on the body of evidence is not fully captured by counting agreements, contradictions and judgment suspensions regarding the final consensus.[28] Some evidential claims seem to have more impact on this certain ratio than others.

Second, there are differences between confirmation measures. Results are the same for both relevance confirmation measures, the only exceptions being De la Beche and Lyell at time step $5a$ and $6a$, respectively. In both cases, the success ratio is greater 0.5, only if using $Z_{DOJ}(h, e)$. Most of the time, $DOJ(h|e)$ and $Z_{DOJ}(h, e)$ agree on persons but differ in the actual value of the ratio. For the same person, the value of the ratio is most of the time smaller using $DOJ(h|e)$. There are also time steps and persons, where the success ratio is greater 0.5, only if not using $DOJ(h|e)$. So in order to maximize the success ratio, using $Z_{DOJ}(h, e)$ is better than using one of the other confirmation measures.

Third, the success ratio depends on time. This is true for all three confirmation measures. For all three confirmation measures and persons, before time step $4a$, it holds that the ratio equals or is less than 0.5. From time step $7b$ until the

[28] Is the success ratio's dependence on the body of evidence better captured by only counting agreements and contradictions regarding the final consensus? As a result of further analyses, for all three confirmation measures, it does not hold that the success ratio is greater 0.5, iff the number of contradictions and agreements are sufficiently small and large, respectively, see sec. E.2.3.

end, all three confirmation measures agree on Murchison, Lyell, Phillips and Austin exhibiting a success ratio of 1.[29] At the final step, this is true for all persons. This result illustrates very nicely the so-called compromise model as introduced in (Kitcher, 1993, p. 201) which models the closure of major scientific debates and embodies ideas of rationalism as well as anti-rationalism:

> "($C1$) The community decision is reached when sufficiently many sufficiently powerful subgroups within the community have arrived at decisions [...] to modify their practices in a particular way.
>
> ($C2$) Scientists are typically moved by non-epistemic as well as epistemic goals.
>
> ($C3$) There is a significant cognitive variation within scientific communities, in terms of individual practices, underlying propensities, and exposure to stimuli.
>
> ($C4$) During early phases of scientific debate, the processes undergone by the ultimate victors are (usually) no more well designed for promoting cognitive progress than those undergone by the ultimate losers.
>
> ($C5$) Scientific debates are closed when, as a result of conversations among peers and encounters with nature that are partially produced by early decisions to modify individual practices, there emerges in the community a widely applicable argument, encapsulating a process for modifying practice, which, when judged by ES [...], is markedly superior in promoting cognitive progress than other processes undergone by protagonists in the debate; power accrues to the victorious group principally in virtue of the integration of this process into the thinking of the members of the community and recognition of its virtue."

Here, ES is some epistemic standard defined as follows in (Kitcher, 1993, p. 189):

> "(ES) The shift from one individual practice to another was rational iff the process through which the shift was made has a success ratio at least as high as that of any other process used by human beings (past, present and future) across the set of epistemic contexts that includes all possible initial practices (for human beings) and possible stimuli (given the world as it is and the characteristics of the human recipient)."

[29] Remember Lyell being discarded from further analyses at time step $8a$ due to his dialectically inconsistent body of evidence.

In (Kitcher, 1993), the great Devonian controversy serves as an illustrative example for the compromise model. After having read (Rudwick, 1988), it is clear that conditions $C1$, $C2$ and $C3$ are met. Answering the question, if conditions $C4$ and $C5$ are met, needs some further analyses. As Kitcher (1993) puts it: "issues around $C4$ and $C5$ are more murky". My analyses sheds new light on these very issues around $C4$ and $C5$.

Some words on meeting $C4$. During early phases of the debate, that is from the beginning until time step $S4$, maximizing evidential support is not well designed to approach the final dating hypothesis, independent of a certain person. Presume that participants of the great Devonian controversy undergo the process of maximizing evidential support. As my analyses show this is true at least most of the time, compare Tab. 3.3 and Fig. 3.16. Presume further, that cognitive progress is considered as approaching the final dating hypothesis. Then, during early phases of the debate, for an arbitrary participant, the process she undergoes is not very well designed for promoting cognitive progress. Hence, $C4$ is met.

Some words on meeting $C5$. During late phases of the great Devonian controversy, that is from time step $7b$ until the end, maximizing evidential support is maximally well designed to approach the final dating hypothesis. However, this result is, at least until the penultimate time step, not independent of a certain person. From time step $7b$ until the end, Murchison, Lyell, Phillips and Austen exhibit a success ratio which equals 1. At the final step, this is true for all persons. Hence, the debate is closed when, as a result of argumentation and evidence accumulation, maximizing evidential support is maximally well designed in approaching the final dating hypothesis, independent of a certain person. Presume once again that (i) participants of the great Devonian undergo the process of maximizing evidential support and (ii) cognitive progress is considered as approaching the final dating hypothesis. Then, the debate is closed when, as a result of argumentation, evidence accumulation and belief changes, for an arbitrary participant, the process she undergoes is maximally well designed for promoting cognitive progress. Hence, $C5$ is met.

ALL			
	DOJ	Z	F
0b	–	–	–
1a	–	–	–
1b	–	–	–
2a	–	–	–
2b	–	–	–
3a	–	–	–
3b	–	–	–
4a	–	((DLB, 0.56))	((DLB, 0.56))
4b	((PHI, 0.7))	((PHI, 1.))	((PHI, 1.))
5a	((MUR, 0.61), (PHI, 0.73))	((DLB, 1.), (MUR, 0.69), (LYE, 1.), (PHI, 1.))	((MUR, 0.69), (LYE, 1.), (PHI, 1.))
5b	((MUR, 0.51), (PHI, 0.7))	((MUR, 0.89), (PHI, 1.))	((MUR, 0.89), (PHI, 1.))
6a	((MUR, 0.51), (PHI, 0.7))	((MUR, 0.89), (LYE, 1.), (PHI, 1.))	((MUR, 0.89), (PHI, 1.))
6b	((LYE, 0.67), (PHI, 0.7), (AUS, 0.73))	((LYE, 1.), (PHI, 1.), (AUS, 1.))	((LYE, 1.), (PHI, 1.), (AUS, 1.))
7a	((LYE, 0.67), (PHI, 0.7), (AUS, 0.73))	((LYE, 1.), (PHI, 1.), (AUS, 1.))	((LYE, 1.), (PHI, 1.), (AUS, 1.))
7b	((MUR, 1.), (LYE, 1.), (PHI, 1.), (AUS, 1.))	((MUR, 1.), (LYE, 1.), (PHI, 1.), (AUS, 1.))	((MUR, 1.), (LYE, 1.), (PHI, 1.), (AUS, 1.))
8a	((MUR, 1.), (PHI, 1.), (AUS, 1.))	((MUR, 1.), (PHI, 1.), (AUS, 1.))	((MUR, 1.), (PHI, 1.), (AUS, 1.))
8b	((DLB, 1.), (MUR, 1.), (LYE, 1.), (PHI, 1.), (SED, 1.), (AUS, 1.))	((DLB, 1.), (MUR, 1.), (LYE, 1.), (PHI, 1.), (SED, 1.), (AUS, 1.))	((DLB, 1.), (MUR, 1.), (LYE, 1.), (PHI, 1.), (SED, 1.), (AUS, 1.))

Figure 3.25.: Ratio of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize evidential support. Only those values are shown which are greater than 0.5.

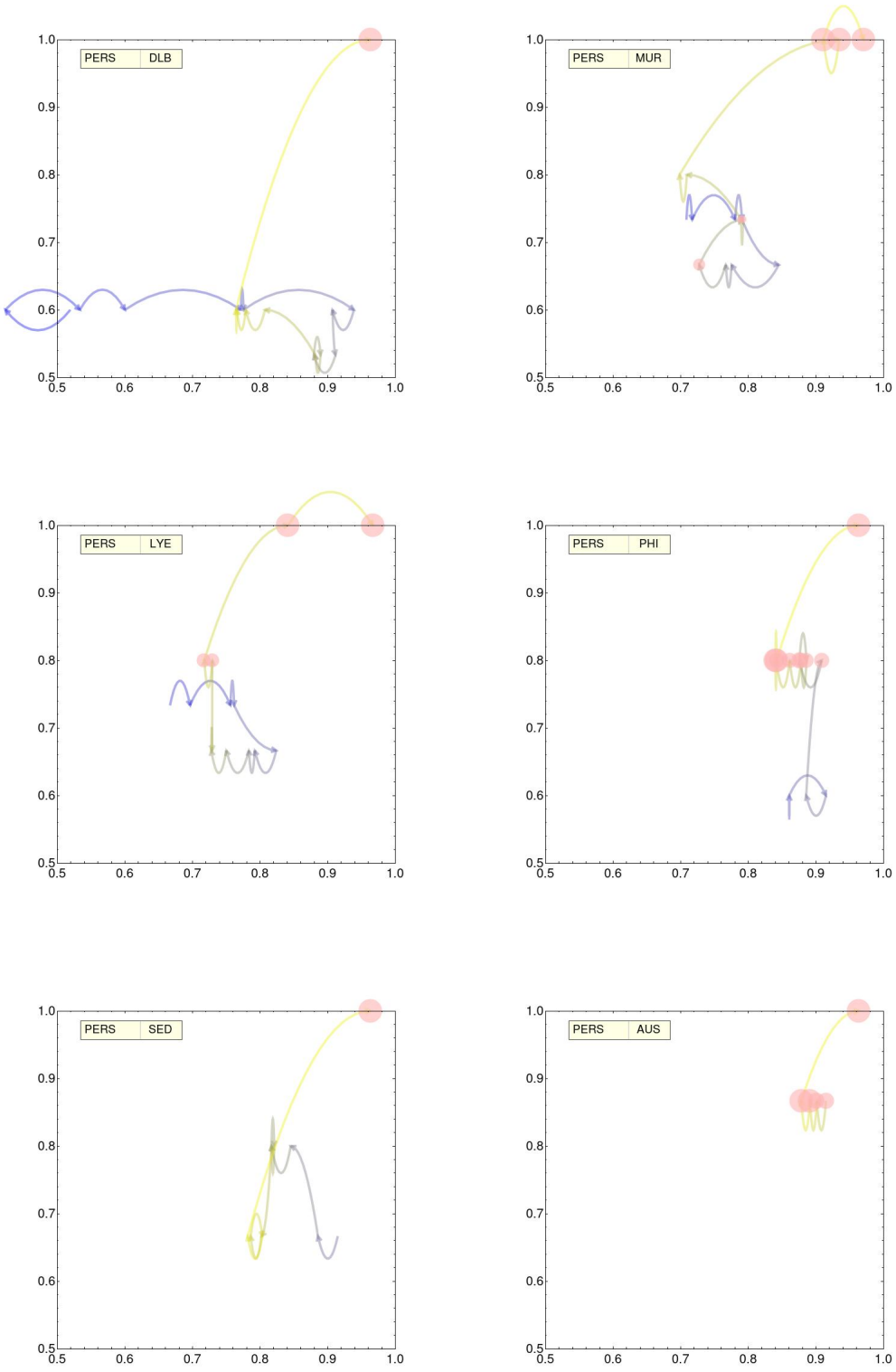

Figure 3.26.: Similarity with the final consensus is shown as in Fig. 3.17. Additionally, those time steps are marked with a red circle where the relative frequency of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize $DOJ(h|e)$ is greater than 0.5. The size of the red circle relates to the actual value of this ratio.

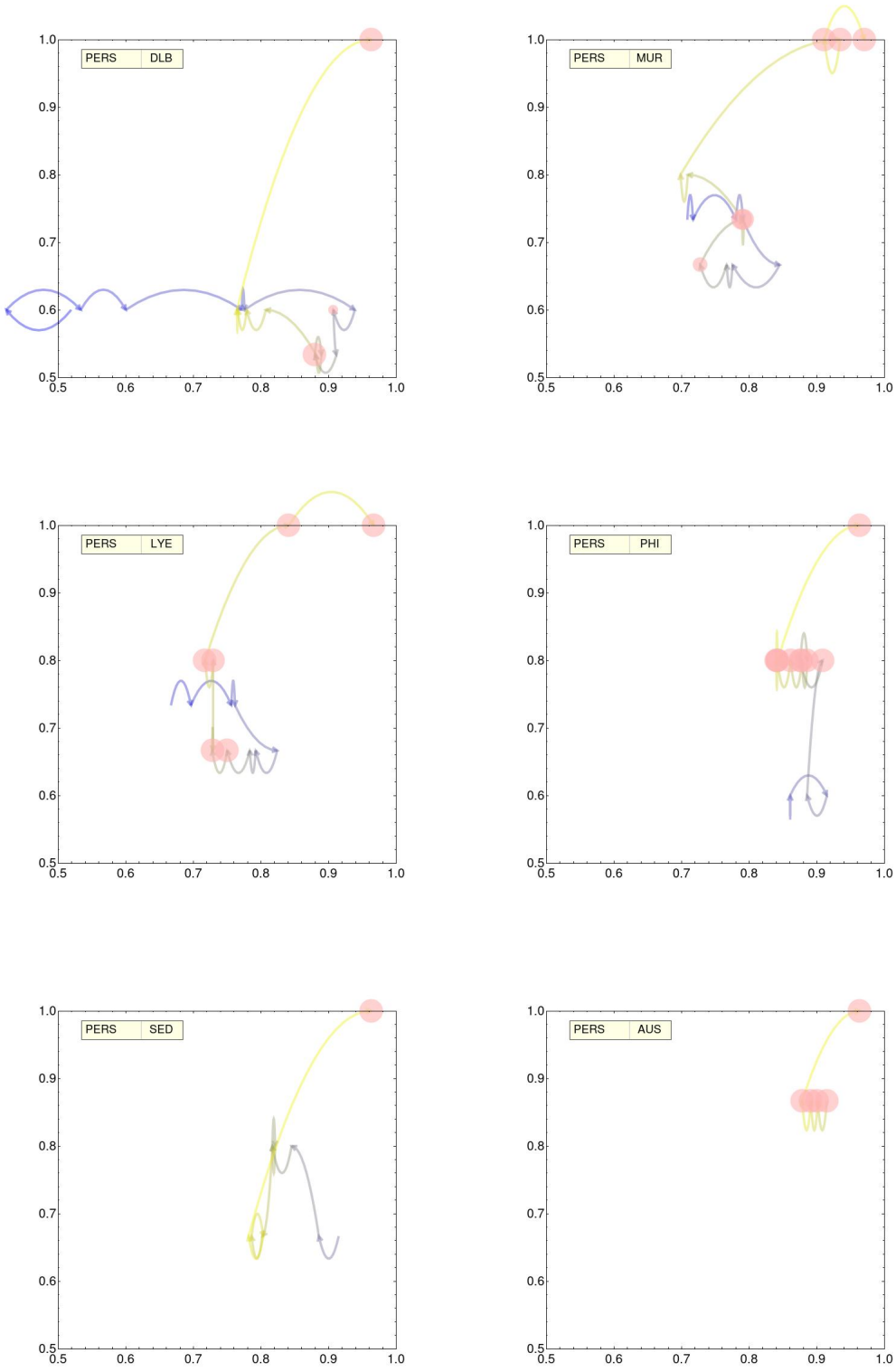

Figure 3.27.: Similarity with the final consensus is shown as in Fig. 3.17. Additionally, those time steps are marked with a red circle where the relative frequency of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize $Z_{DOJ}(h, e)$ is greater than 0.5. The size of the red circle relates to the actual value of this ratio.

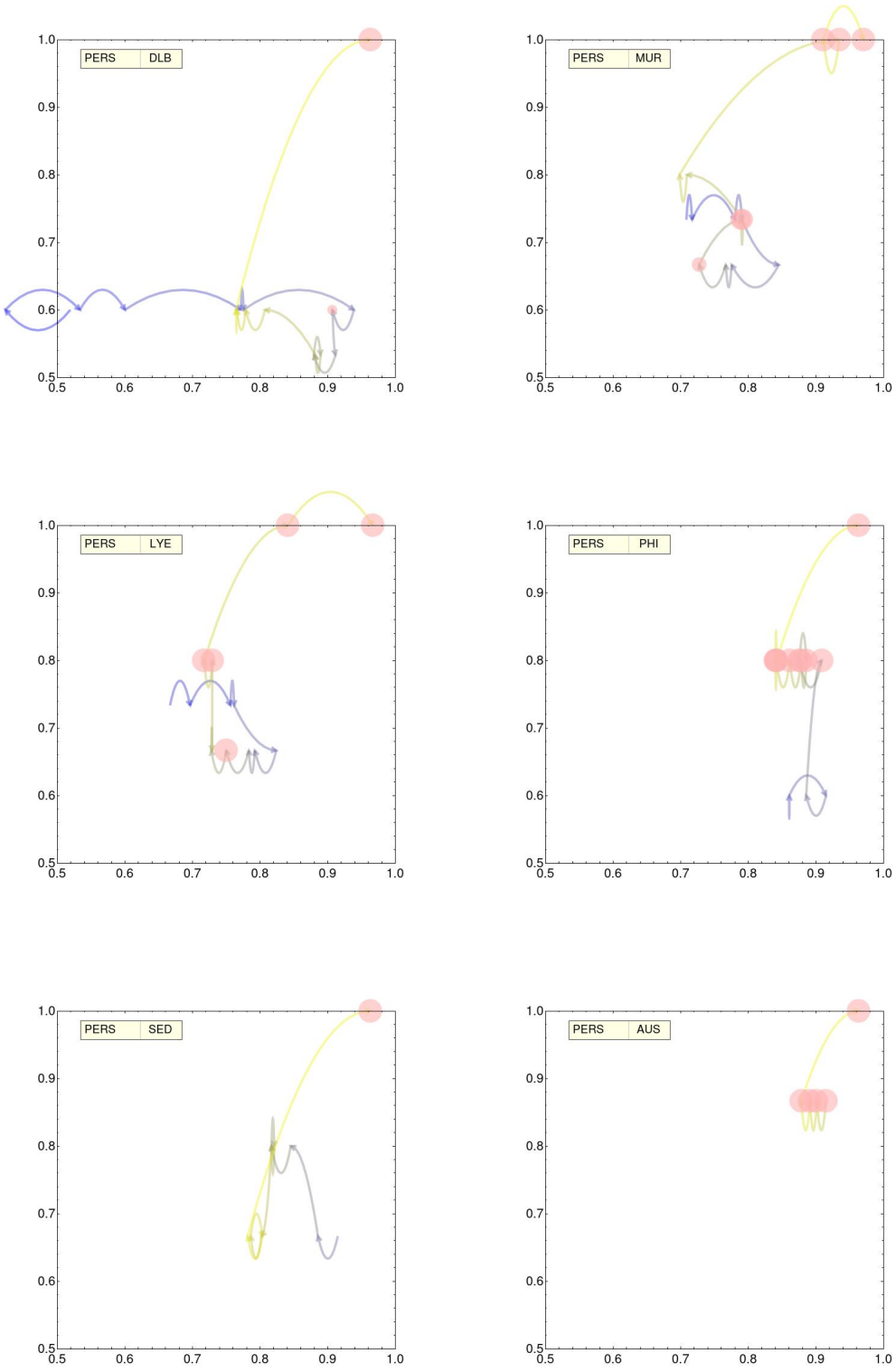

Figure 3.28.: Similarity with the final consensus is shown as in Fig. 3.17. Additionally, those time steps are marked with a red circle where the relative frequency of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize $F_{DOJ}(h, e)$ is greater than 0.5. The size of the red circle relates to the actual value of this ratio.

4. Conclusion

How can we reason with false evidence and reliably infer hypotheses? This is the main question of my thesis which triggers at least two methodological turns, namely a statistical and a historic one. See sec. 1.1 for motivation and points of reference for both methodological turns. For every methodological turn, the main question is further differentiated, resulting in 4 and 16 research questions answered in detail in part one and two of my thesis, that is chapter 2 and chapter 3, respectively. See sec. 1.2 for a detailed listing of these questions. Analyses are performed for three different confirmation measures, namely $DOJ(h|e)$, $Z_{DOJ}(h, e)$ and $F_{DOJ}(h, e)$. The first one assesses absolute confirmation, the others relevance confirmation.

To answer research questions of part one of my thesis, that is chapter 2, 1000 simulated debates are drawn from (Betz, 2013) and analyzed in terms of confirmation. Analyses focus on three examples of higher-order evidence, namely the inferential density of the dialectical structure at a certain time step, and, for every person, the amount of evidence claims and the ratio of true evidence claims. The analyses comprise the following steps:

1. For every debate, inferential density is calculated, truth is set and an evidence stream is generated. As in (Betz, 2015), truth setting as well as evidence stream generation makes use of random numbers. An evidence stream is an accumulative list of lists of evidential beliefs. In contrast to (Betz, 2015), a certain ratio of these evidential beliefs are false.

2. For every hypothesis of a certain debate, the degree of confirmation is calculated given (*i*) a certain amount of evidential beliefs, (*ii*) a certain ratio of true evidential beliefs and (*iii*) a certain inferential density.

3. A statistical hypothesis test is performed with confirmation as test statistics. The hypothesis $\neg h$ is tested against the alternative hypothesis h.

    a) The critical region is chosen such that the power of the test equals 0.25.

b) The significance of the test is calculated.

The main findings of the analyses are:

- ($V1$) The relative frequency of truths among hypotheses with a certain degree of confirmation. It shows that, for a totally correct body of evidence, higher-order evidence influences relative frequency, only if using a relevance confirmation measure.
- ($V2$) The reliability of confirmation as a veritistic indicator. Independent of higher-order evidence, it shows that absolute confirmation is a more reliable indicator than relevance confirmation. Higher-order evidence influences the reliability of confirmation as a veritistic indicator and there are differences between absolute and relevance confirmation.
  - ($V2.0$) As the ratio of false evidence claims increases, the reliability decreases.
  - ($V2.1$) As the inferential density increases, the reliability increases. For a body of evidence including false evidence claims and absolute confirmation, the truth of $V2.1$ depends on the amount of evidence.
  - ($V2.2$) As more and more evidence is accumulated, the reliability increases. For a body of evidence including false evidence claims, the truth of $V2.2$ depends on the inferential density. However, there are exceptions. First, for a body of evidence including false evidence claims and a certain kind of relevance confirmation, $V2.2$ holds, independent of the inferential density. Second, for a sufficiently large amount of false evidence claims and absolute confirmation, $V2.2$ does not hold, independent of the inferential density.
  - ($V2.3$) As more and more evidence is accumulated, differences in reliability between absolute and relevance confirmation decrease.
  - ($V2.4$) As the inferential density increases, differences in reliability between absolute and relevance confirmation decrease.
- ($V3$) Situations in which its degree of confirmation is evidence for the confirmed hypothesis. There are situations where its degree of confirmation is a reliable indicator for the truth of a hypothesis. It shows that absolute confirmation is more often a reliable veritistic indicator than relevance confirmation.

For a sufficiently small amount of true evidence claims, confirmation is no reliable veritistic indicator, independent of a certain confirmation measure.

- ($V4$) Exceptions to the epistemic rule of forming beliefs according to confirmation. There are situations where it is not rational to form beliefs according to confirmation, namely those where confirmation is no reliable indicator for the truth of a hypothesis.

Some remarks on $V1$. Presuppose subjects respond to a hypothesis in accordance with its degree of confirmation given some body of evidence. In what circumstances is this process reliable? For a totally true body of evidence, Betz (2015) has shown that the inferential density as well as amount of evidence allows us to estimate the reliability of absolute confirmation as a veritistic indicator for the truth of a hypothesis. Betz (2015) assesses reliability in terms of the relative frequency of truths among hypotheses with a certain degree of confirmation. Hence, $V1$ expands upon (Betz, 2015) by considering (i) partly incorrect bodies of evidence and (ii) relevance confirmation measures. Further, there are several re-evaluating principles, stating that I have to re-evaluate my former beliefs in light of certain evidence about the process producing these beliefs. $V1$ underpins a certain re-evaluating principle, namely the $MERF$ principle as introduced in (Talbott, 2016b), in a twofold way. First, it assesses the reliability of a special cognitive process, namely belief formation according to confirmation. Second, it identifies reliability-relevant categorizations of this very process in terms of higher-order evidence, namely the inferential density of the dialectical structure, the amount of evidence and the ration of true evidence claims.

Some remarks on $V2$. $V2$ expands upon (Betz, 2015) not only by considering (i) partly incorrect bodies of evidence and (ii) relevance confirmation measures but also by assessing the reliability of confirmation as a veritistic indicator via significance and power of a corresponding statistical hypothesis test. $V2$ underpins a certain re-evaluating principle, namely the integration principle as introduced in (Christensen, 2008), in a twofold way. First, it assesses the reliability of a special cognitive process, namely belief formation according to a statistical hypothesis test with confirmation as test statistics. Second, it identifies reliability-relevant categorizations of this very process in terms of higher-order evidence, namely the inferential density of the dialectical structure, the amount of evidence and the ration of true evidence claims.

Some remarks on $V3$. Let us suppose that confirmation is evidence of evidence.

Then, $V3$ shows that there are situations, in which evidence of evidence (for some hypothesis h) is itself evidence for h, with a situation being characterized by some other examples of higher-order evidence, namely the inferential density of the dialectical structure, the amount of evidence and the ration of true evidence claims. This thesis is claimed by some epistemologists, see for example Feldman (2005). However, it is also contested by some others, see for example Fitelson (2012).

Some remarks on $V4$. For all three confirmation measures, there are situations where confirmation is no reliable veritistic indicator, with a situation being characterized by higher-order evidence, namely the inferential density of the dialectical structure, the amount of evidence and the ration of true evidence claims. Therefore, adjusting one's belief according to confirmation is not always rational. Presume that adjusting one's beliefs according to confirmation is an epistemic rule. Then, my analysis shows that there are exceptions to this rule. Hence, it confirms Feyerabend (1976) stating that there are no strict epistemic rules.

To answer research questions of part two of my thesis, that is chapter 3, time span of an historic debate, namely the great Devonian controversy, is discretized and the following is performed for every time step:

1. The dialectical structure is reconstructed, identifying auxiliary assumptions, inferential relations, argument types and clusters.
2. For every main participant and the final consensus, all evidential beliefs and a dating hypothesis are identified.
3. Groups of persons are defined exogenously as well as endogenously using a newly introduced similarity measure.
4. For every main participant, her dating hypothesis' degree of confirmation given all her evidential beliefs is calculated.
5. For every main participant, similarity with the final consensus is assessed using the previously introduced similarity measure.
6. For every main participant, the relative frequency of dating hypotheses which are sufficiently similar with the final dating hypothesis among all those dating hypotheses maximizing evidential support is determined.

These analyses foster understanding and deepen knowledge of several important philosophical concepts and issues:

- ($H1$) Relations between evidence and hypotheses

- ($H2$) Consensus and consensus dynamics
- ($H3$) Individual belief changes
- ($H4$) Rational belief change
- ($H5$) Consensus formation

Some words on $H1$, that is relations between evidence and hypotheses. Reconstructing the great Devonian controversy, the following shows:

- ($H1.1$) For all empirical statements of the great Devonian controversy, it holds: There is a dependence between its theoretical context and rational acceptance. Hence, the debate illustrates nicely the concept of theory-ladenness, which is uncontroversial in today's philosophy of science, see for example (Boyd and Bogen, 2021).
- ($H1.2$) For the great Devonian controversy, it shows that an empirical statement is implied by some mineralogical or fossil criterion, only if it is conjoined with some auxiliary assumptions. Therefore, the debate illustrates Duhemian underdetermination, compare (Duhem, 1954).
- ($H1.3$) There are several criteria and most of them are highly controversial most of the time. Only at the end, there is a criterion which all participants agree upon. Therefore, the great Devonian controversy illustrates the struggle about standardizing methodological rules for generating empirical statements.

Not only inferential relations between evidence and dating hypotheses are revealed, but also the notion of confirmation is quantified. Three different confirmation measures are compared, namely $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. Calculating degrees of confirmation, the following shows.

- ($H1.4$) The great Devonian controversy starts and ends with all main participants accepting a dating hypothesis with a maximal degree of confirmation (given a certain evidence), that is with a degree of 1, independent of a certain confirmation measure.
- ($H1.5$) For most time steps and participants, $DOJ(h|e)$ and $Z_{DOJ}(h,e)$ are rather similar, both in value and relative changes, and much smaller than 1.
- ($H1.6$) For most time steps and participants, $DOJ(h|e)$ and $F_{DOJ}(h,e)$ are rather unsimilar, both in value and relative changes. For most time steps and participants, $F_{DOJ}(h,e)$ increases with increasing $\frac{DOJ(h|e)}{DOJ(h)}$ and is fairly approximated by 1.

Some words on $H2$, that is consensus and consensus dynamics. For every time step, not only groups are identified exogenously by accepting a certain piece of evidence, but endogenously using degrees of similarity between dating hypotheses respectively bodies of evidence. This way, (Rudwick, 1988) is enhanced. The main findings of these analyses are:

- ($H2.1$) For exogenous clustering, groups are never the same as those in (Rudwick, 1988). This is not true for endogenous clustering. Clustering maximally similar dating hypotheses, groups are the same as those in (Rudwick, 1988).
- ($H2.2$) Similarity spectra of dating hypotheses and bodies of evidence are quite similar, namely $[0.60, 1.00]$ and $[0.54, 0.99]$. However, similarity dynamics of dating hypotheses and bodies of evidence do not coincide:
  - Except during the middle section, there are always some persons accepting the same dating hypothesis. Never, not even at the end, there are two persons accepting the same body of evidence.
  - Several times, there are two persons accepting, at the same time, remarkably similar bodies of evidence but unsimilar dating hypotheses, and vice versa.
- ($H2.3$) The average degree of similarity is maximal at the final step, not at last due to argumentation. In so far as dating hypotheses and bodies of evidence together constitute a paradigm, this result may be understood as an illustration of Kuhn (1983) stating that controversies are not only triggered but also resolved by inter-paradigmatic exchange of arguments.

Some words on $H3$, that is individual belief changes. Analyzing individual belief changes of participants of the great Devonian controversy, the following shows:

- ($H3.1$) Participants do not only change their dating hypotheses, but also their evidential beliefs. Often, participants hold on to a certain dating hypothesis while changing evidential beliefs. Given that participants are rational, this result dis-confirms strict falsificationism in the sense of Popper (1935).
- ($H3.2$) For the great Devonian controversy, there are dating hypotheses as well as evidential beliefs, which are constantly kept, or at least only very reluctantly given up. Hence, these dating hypotheses and evidential beliefs can be considered as hard core assumptions in the sense of Lakatos (1970).

- ($H3.3$) Most of the time, dating hypotheses as well as bodies of evidence are only slightly altered. This illustrates Laudan (1984) stating that beliefs are not revised as a whole, but rather in a piecemeal and reluctant way.

Some words on $H4$, that is rational belief change. Is rationality in belief change related with some kind of evidential support? Here, evidential support is spelled out in terms of confirmation, that is using $DOJ(h|e)$, $Z_{DOJ}(h,e)$ or $F_{DOJ}(h,e)$. The following principles of rational belief change are tested:

- ($RAT1$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it does not decrease the dating hypothesis's degree of evidential support.
- ($RAT2$) Shifting from one group of a dating hypothesis and evidential beliefs to another one is only rational, if it maximizes the dating hypothesis's degree of evidential support.

Here, a person maximizes her dating hypothesis's evidential support, if she chooses a dating hypothesis such that, given her body of evidence, there is no better confirmed dating hypothesis. For the great Devonian controversy, the following shows:

- ($H4.1$) Most of the time, individual belief changes are rational. However, there are individual belief changes which are not rational. Using $F_{DOJ}(h,e)$, individual belief changes are *more* often rational than using one of the other two confirmation measures.
- ($H4.2$) Shared belief changes are *less* often rational than individual belief changes. Using $F_{DOJ}(h,e)$, shared belief changes are *more* often rational than using one of the other two confirmation measures.

Presuppose that participants of the great Devonian controversy are rational. Together with $H4.1$ it follows that there are exceptions to the two previously introduced principles of rational belief change. This confirms Feyerabend (1976) stating that there is no scientific rule without any exceptions. However, it does not support relativism in the sense of "anything goes".

Some words on $H5$, that is consensus formation. Roads to the final consensus differ remarkably. There are no two persons following the same road. However, there are rather strong similarities. For all main participants of the great Devonian controversy, it holds:

- ($H5.1$) Approaching alternates with distancing the final consensus. This is in line with Betz (2013) investigating consensus-conduciveness of controversial debates by means of multi-agent simulations.
- ($H5.2$) Approaching the final consensus in terms of dating hypotheses does not imply an approachment in terms of bodies of evidence, and vice versa. This is a refinement of the analysis of consensus formation given in (Rudwick, 1988).

As a result of my analyses, it does not hold that a belief change distances the final consensus, iff it decreases evidential support. This is true for all three confirmation measures. However, it shows that, after a sufficiently large number of successive belief changes decreasing evidential support, there is often a considerable change in similarity with the final consensus. Hence, successive decrease of evidential support seems to be a reason for changing beliefs. Note that there are reasons for not changing beliefs, even if evidential support decreases.

As a further result of my analyses, it does not hold that a belief change distances the final consensus, iff it does not maximize evidential support. This is true for all three confirmation measures. Further, it does not hold that a belief change approaches the final consensus, if it maximizes evidential support.

Is maximizing evidential support well designed to approach the final consensus, that is, do most dating hypotheses maximizing evidential support have a sufficiently high degree of similarity with the final dating hypothesis? For every time step and person, the relative frequency of dating hypotheses which are sufficiently similar with the final dating hypothesis among all those dating hypotheses maximizing evidential support is determined. There are several things about this ratio which should be noted:

- ($H5.3$) There are big differences between persons. Hence, there is a dependence between this ratio and a person's body of evidence. However, it cannot be assessed in terms of a body's similarity with the final consensus. A certain degree of similarity is neither a sufficient nor a necessary condition for the ratio being greater 0.5. Hence, some evidential claims seem to have more impact than others.
- ($H5.4$) There are differences between confirmation measures. In order to maximize the ratio, $Z_{DOJ}(h, e)$ is better than the other two confirmation measures.

- ($H5.5$) For all three confirmation measures and persons, before time step $4a$, it holds: The ratio is less or equal 0.5. Hence, during early phases of the debate, maximizing evidential support is not well designed to approach the final consensus, independent of a certain person.
- ($H5.6$) For all three confirmation measures and most main participants, from time step $7b$ till the end, the ratio equals 1. At the final step, this is true for all persons. Hence, the debate is closed when, as a result of argumentation, evidence accumulation and belief changes, maximizing evidential support is maximally well designed to approach the final dating hypothesis, independent of a certain person.

The two latter results newly illustrate a model of the closure of major scientific debates embodying ideas of rationalism as well as anti-rationalism, namely the compromise model as introduced in (Kitcher, 1993). Presume that (i) participants of the great Devonian controversy undergo the process of maximizing evidential support and (ii) cognitive progress is considered as approaching the final dating hypothesis. Then, $H5.5$ confirms condition $C4$ of the compromise model stating that "[d]uring early phases of scientific debates, the processes undergone by the ultimate victors are (usually) no more well designed for promoting cognitive progress than those undergone by the ultimate losers" (Kitcher, 1993, p. 201). Further, presuming the same two assumptions, $H5.6$ confirms condition $C5$ of the compromise model, roughly stating that scientific debates end when, as a result of argumentation, evidence accumulation and belief changes, a certain cognitive process, which is executable for all participants, performs better in terms of cognitive progress than all the others undergone by participants of the debate.

Danksagung

Dank schulde ich stets meiner Familie und meinen Freunden für ihre Unterstützung und ihr Vertrauen. Bedanken möchte ich mich bei Herrn Professor Betz für seine unermüdliche und überaus kompetente Betreuung. Nicht vergessen möchte ich die freundliche Hilfsbereitschaft von Frau Schwarzenberger und den Teilnehmern des Forschungsseminars des philosophischen Instituts.

A. Confirmation as a Veritistic Indicator

A.1. Confirmation Histograms

In this thesis, for every hypothesis of a certain debate, the degree of confirmation is calculated given (*i*) a certain amount of evidential beliefs, (*ii*) a certain ratio of true evidential beliefs and (*iii*) a certain inferential density. Results can be visualized as histograms regarding degrees of confirmation.

In the following, grids of histograms are shown with rows and columns differing in amount of evidence and inferential density, respectively. For every grid, the ratio of true evidential claims is constant. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of confirmation. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

The following three subsections show results for three different confirmation measures, namely $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. The first one assesses absolute confirmation, the others relevance confirmation. For every confirmation measure, the analysis is performed for three different ratios of true evidence claims, namely $1.0, 0.8, 0.6$.

A.1.1. Absolute Confirmation

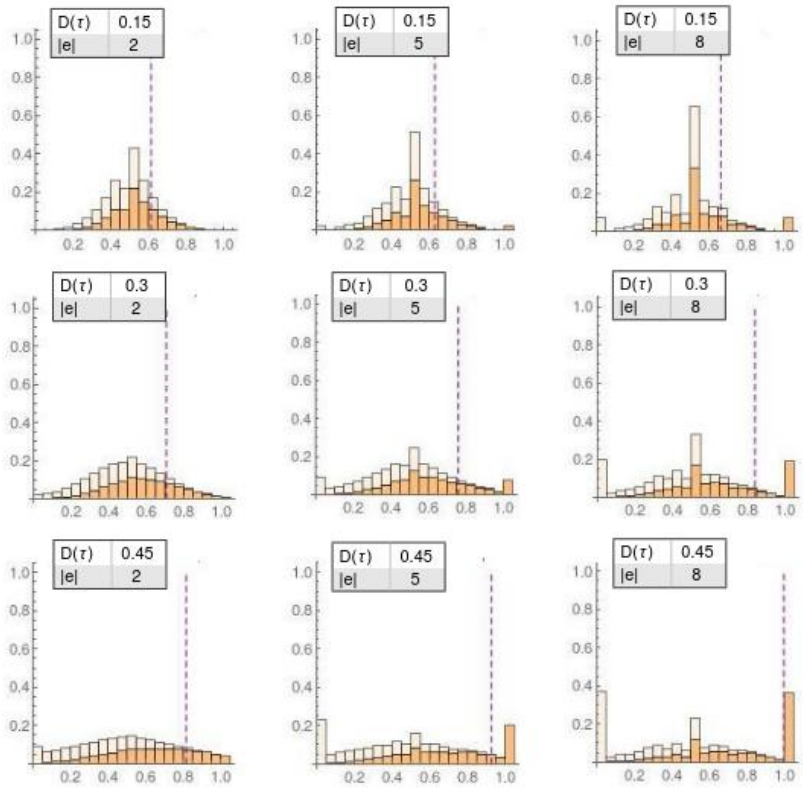

Figure A.1.: A grid of histograms is shown with rows and columns differing in amount of evidence, $|e|$, and inferential density, $D(\tau)$, respectively. The ratio of true evidential claims is constantly 1. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of $DOJ(h|e)$. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

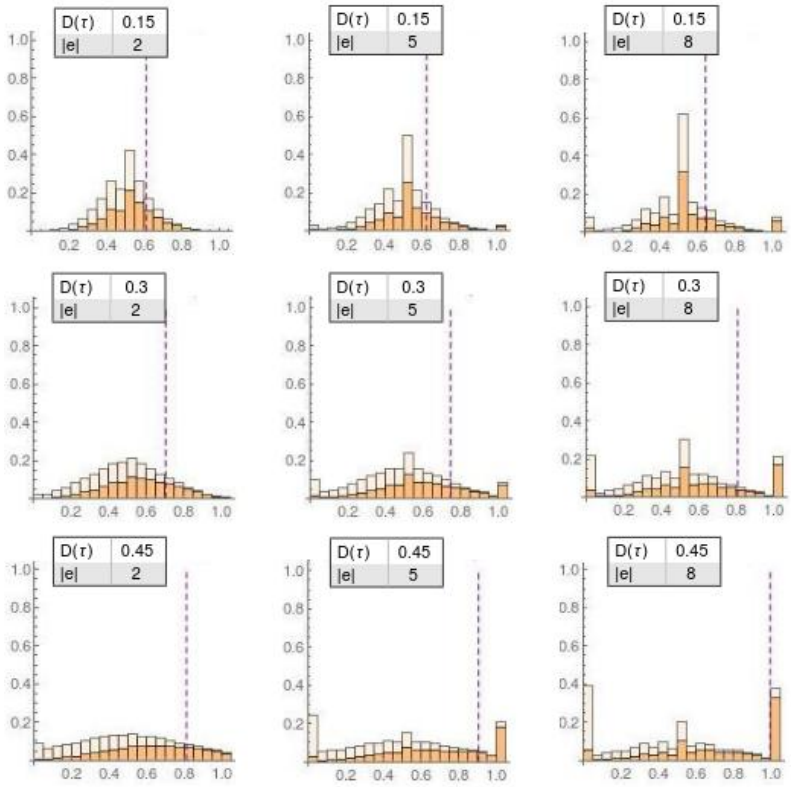

Figure A.2.: A grid of histograms is shown with rows and columns differing in amount of evidence, $|e|$, and inferential density, $D(\tau)$, respectively. The ratio of true evidential claims is constantly 0.8. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of $DOJ(h|e)$. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

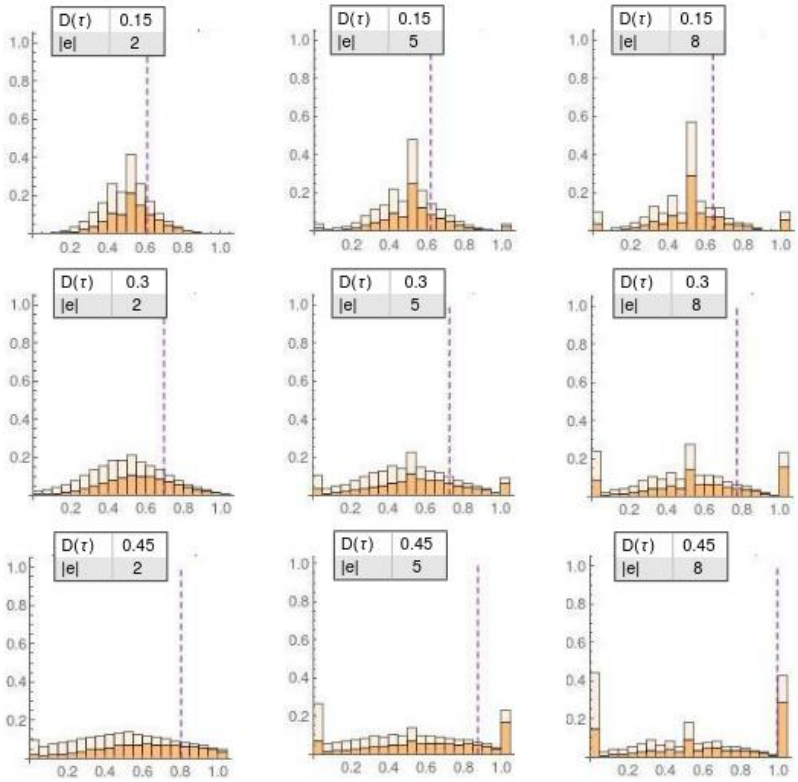

Figure A.3.: A grid of histograms is shown with rows and columns differing in amount of evidence, $|e|$, and inferential density, $D(\tau)$, respectively. The ratio of true evidential claims is constantly 0.6. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of $DOJ(h|e)$. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

A.1.2. Relevance Confirmation I

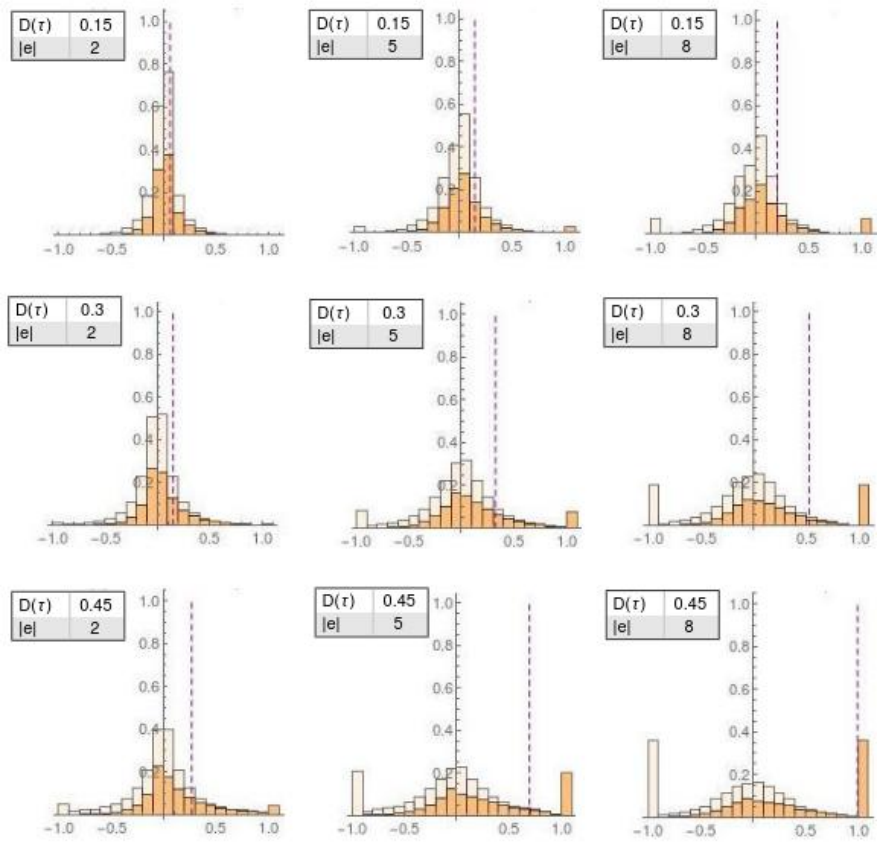

Figure A.4.: A grid of histograms is shown with rows and columns differing in amount of evidence, $|e|$, and inferential density, $D(\tau)$, respectively. The ratio of true evidential claims is constantly 1. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of $Z_{DOJ}(h, e)$. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

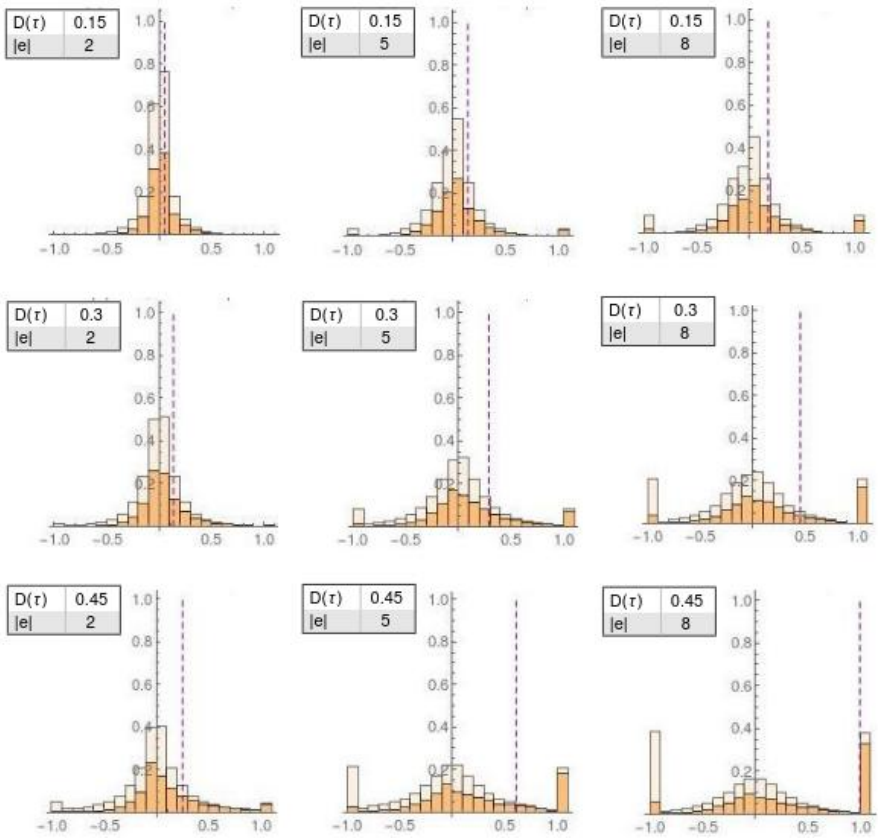

Figure A.5.: A grid of histograms is shown with rows and columns differing in amount of evidence, $|e|$, and inferential density, $D(\tau)$, respectively. The ratio of true evidential claims is constantly 0.8. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of $Z_{DOJ}(h, e)$. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

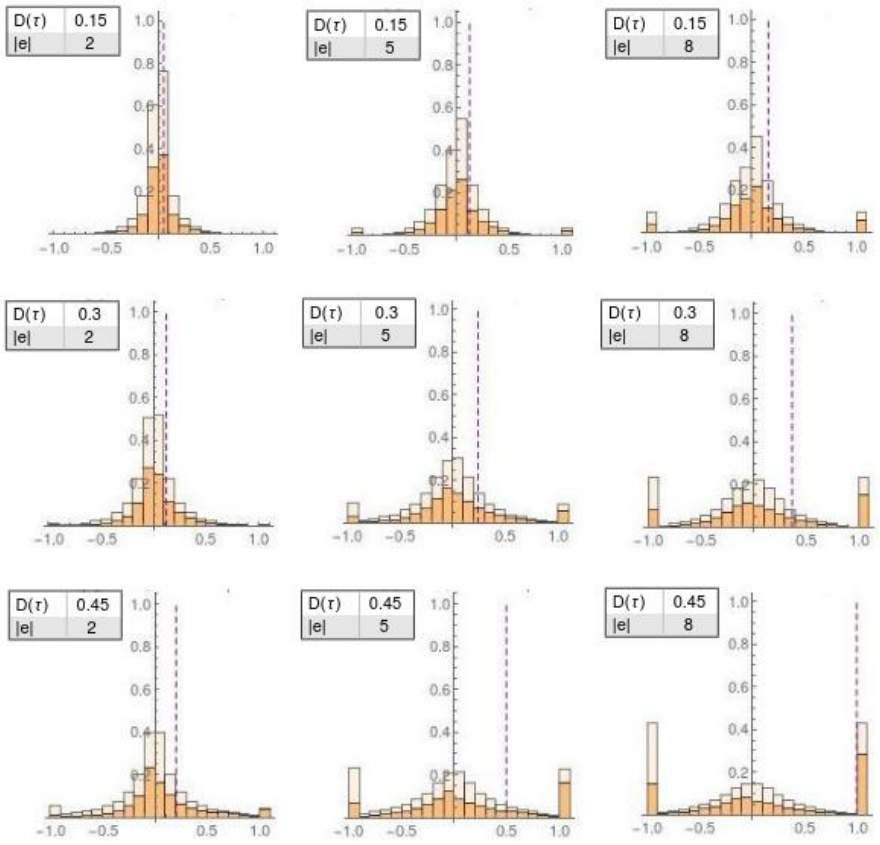

Figure A.6.: A grid of histograms is shown with rows and columns differing in amount of evidence, $|e|$, and inferential density, $D(\tau)$, respectively. The ratio of true evidential claims is constantly 0.6. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of $Z_{DOJ}(h, e)$. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

A.1.3. Relevance Confirmation II

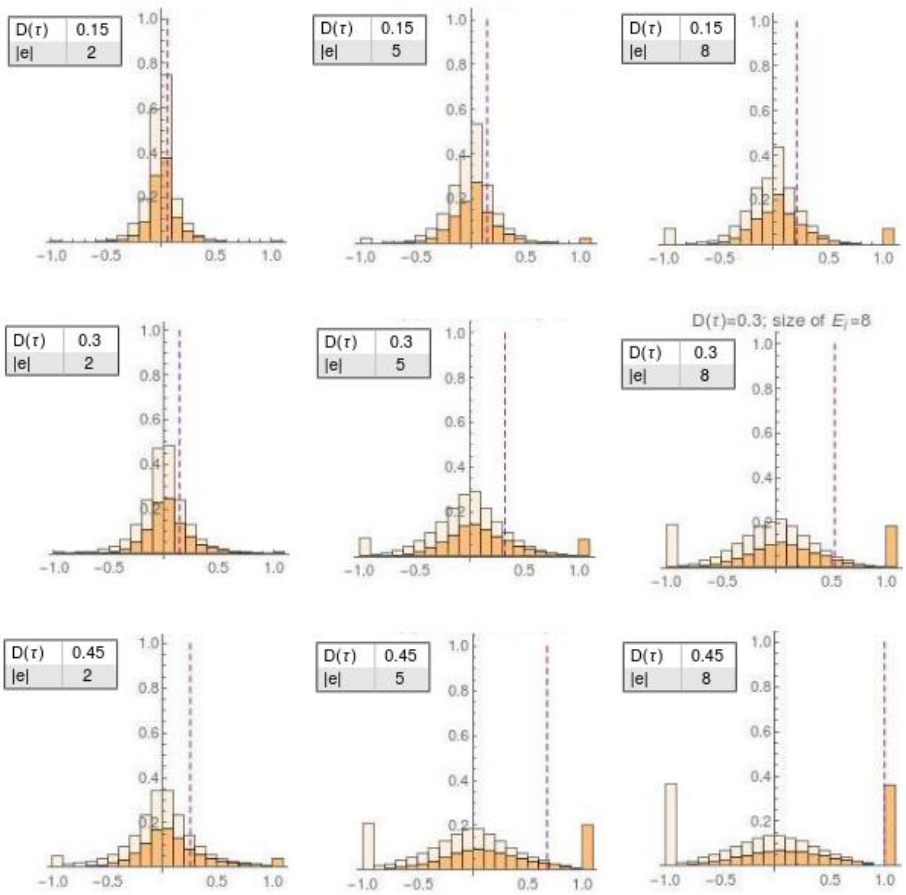

Figure A.7.: A grid of histograms is shown with rows and columns differing in amount of evidence, $|e|$, and inferential density, $D(\tau)$, respectively. The ratio of true evidential claims is constantly 1. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of $F_{DOJ}(h, e)$. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

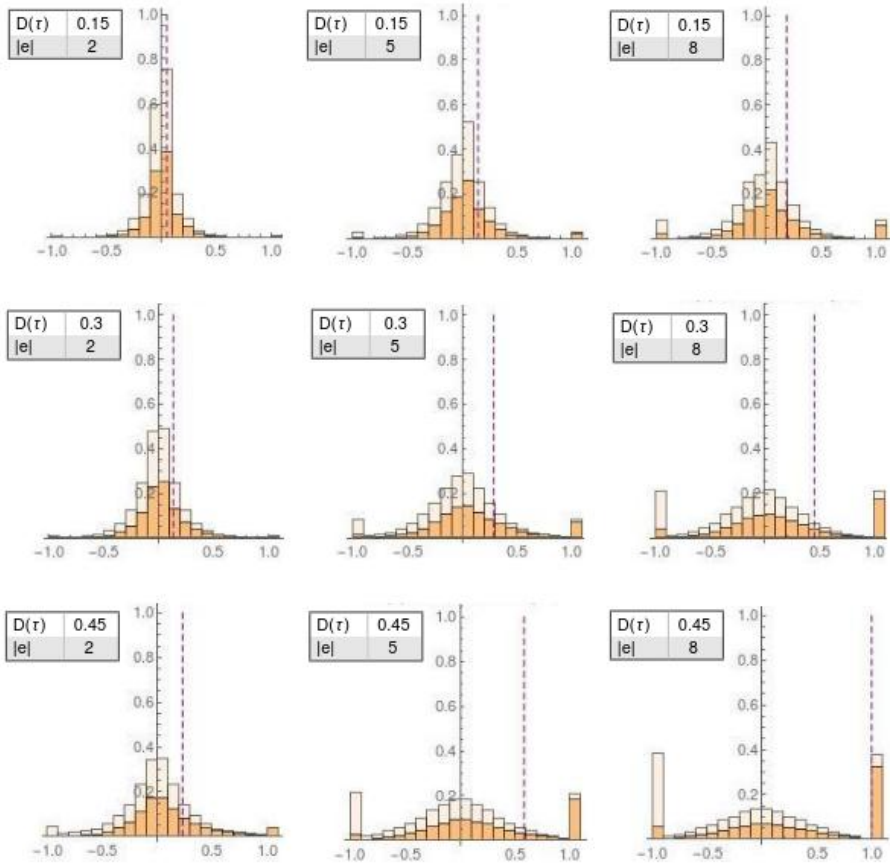

Figure A.8.: A grid of histograms is shown with rows and columns differing in amount of evidence, $|e|$, and inferential density, $D(\tau)$, respectively. The ratio of true evidential claims is constantly 0.8. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of $F_{DOJ}(h, e)$. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

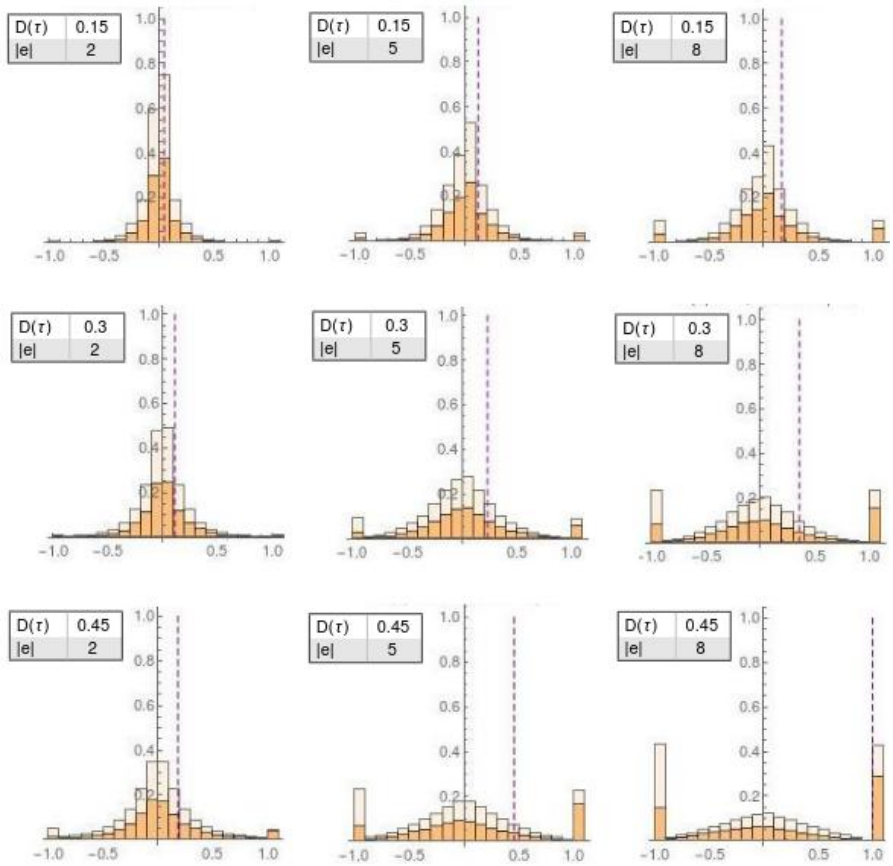

Figure A.9.: A grid of histograms is shown with rows and columns differing in amount of evidence, $|e|$, and inferential density, $D(\tau)$, respectively. The ratio of true evidential claims is constantly 0.6. The bars in darker (lighter) shading represent the fraction of true (false) hypotheses within a certain interval of $F_{DOJ}(h, e)$. Additionally, the critical region of a corresponding hypothesis test with a power of 0.25 is shown, spanning from the dashed line until 1.

A.2. Power of the Statistic Test

For a body of evidence and a hypothesis test with confirmation as the test statistic, what is the power given a significance of 0.05? Does higher-order evidence influence these results?

Fig. A.10 shows β, ranging from 0.43 to 0.94. The power of a statistic hypothesis test is given by $1-\beta$. Independent of the inferential density, the amount of evidence and the ratio of true evidential claims, it shows that the power is largest using $DOJ(h|e)$ and smallest using $F_{DOJ}(h,e)$. Hence, $DOJ(h|e)$ is a more reliable indicator for the truth of a hypothesis than $Z_{DOJ}(h,e)$, and $Z_{DOJ}(h,e)$ is a more reliable indicator for the truth of a hypothesis than $F_{DOJ}(h,e)$.

For a completely correct body of evidence and every confirmation measure, it holds:

- ($R1$) As the inferential density increases, the power increases.
- ($R2$) As more and more evidence is accumulated, the power increases.
- ($R3$) As more and more evidence is accumulated, differences between absolute and relevance confirmation decrease.

Hence, argumentation and evidence accumulation improve the reliability of absolute as well as relevance confirmation as a veritistic indicator. Note that these results confirm those of Betz (2015), using a different way of assessing the reliability of a veritistic indicator.

As the ratio of true evidential claims decreases, things get a little bit more complicated. Only for relevance confirmation, $R1$ holds independent of the amount of evidence. Only for a certain kind of relevance confirmation, $R2$ holds independent of the inferential density and ratio of true evidential claims. For absolute confirmation and a sufficiently small ratio of true evidence claims, $R2$ does not hold independent of the inferential density.

There are situations, in which the power is greater 0.25, see for example Fig. A.11 filtering results according to this very threshold. These situations are most and less often using $DOJ(h|e)$ and $F_{DOJ}(h,e)$, respectively.

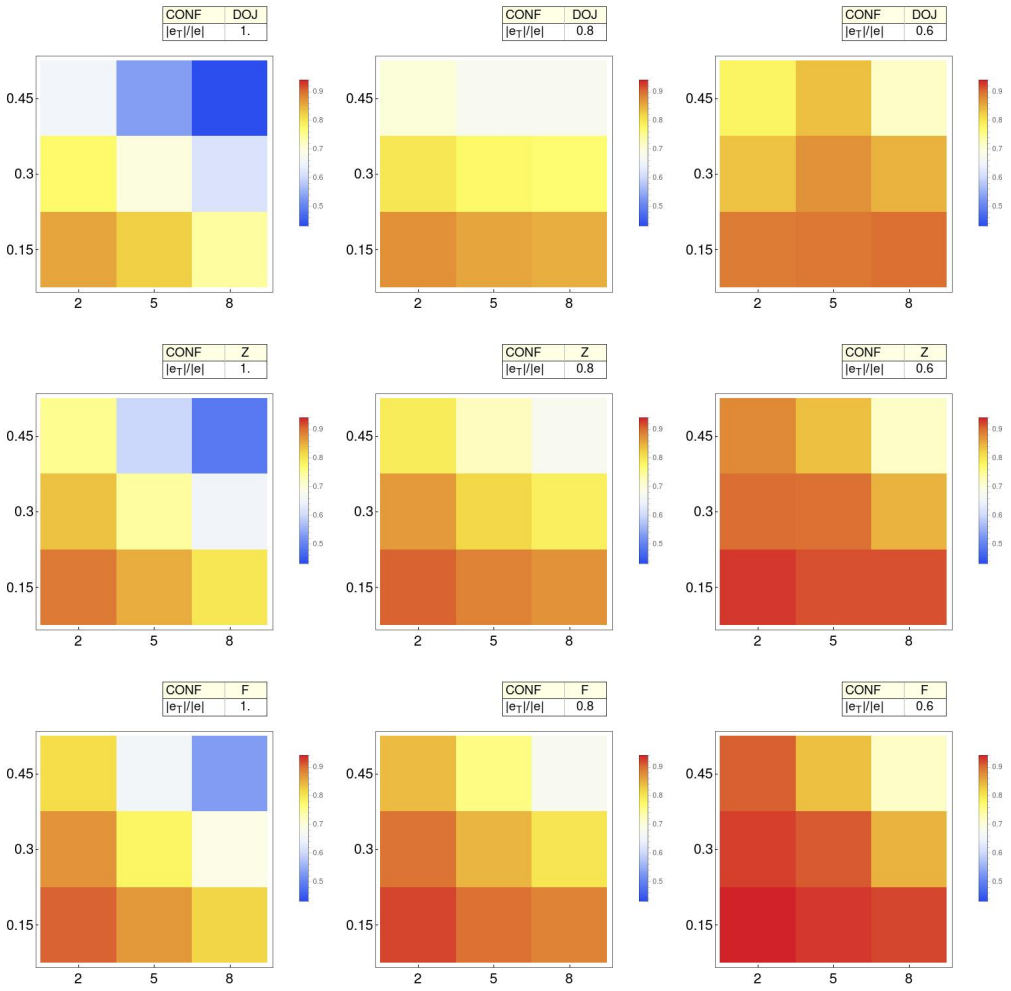

Figure A.10.: β. A grid of plots is shown with rows and columns differing in confirmation measures and ratio of true evidence claims, respectively. From top to bottom, rows correspond to $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. From left to right, the ratio of true evidential claims decreases. Every single plot shows a grid with rows and columns differing in inferential density, and the amount of evidence. For every single plot, from bottom to top, the inferential increases and, from left to right, the amount of evidence increases. β is represented using a color function. Turning from blue over yellow to red, β increases.

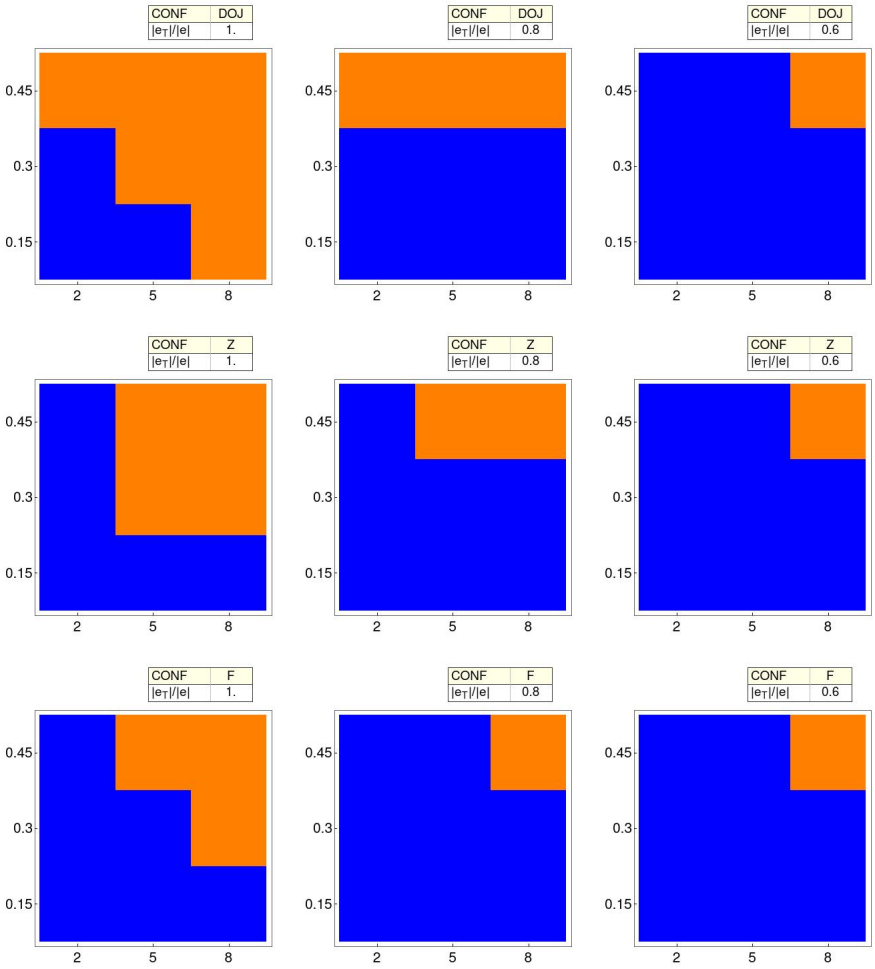

Figure A.11.: $\beta \leq 0.75$. A grid of plots is shown with rows and columns differing in confirmation measures and ratio of true evidence claims, respectively. From top to bottom, rows correspond to $DOJ(h|e)$, $Z_{DOJ}(h, e)$ and $F_{DOJ}(h, e)$. From left to right, the ratio of true evidential claims decreases. Every single plot shows a grid with rows and columns differing in inferential density, and the amount of evidence. For every single plot, from bottom to top, the inferential increases and, from left to right, the amount of evidence increases. Orange and blue refer to $\beta \leq 0.75$ and $\beta > 0.75$, respectively.

B. Reconstruction of the Great Devonian Controversy

B.1. Omissions and Simplifications

Here, omitted sub-debates and sentences are listed:

1. Similarities between several limestones in south Devon. This debate is mainly held by two persons, namely Sedgwick and Austen, who finally reach a consensus on grounds of a sufficiently large amount of fossils.
2. Existence of pseudo-Culm. This idea is supported only by one person, namely De la Beche, for quite a short time.
3. Taxonomic or biological variations and their implications on dating by means of fossils.
4. A limited version of the characteristic rock type dating principle.
5. Observations of rock types from south-west England (Mendip Hills and Manacaan), namely rock types not resembling the one of the main part of the Culm strata, but classified as Coal Measures in age.
6. Existence of a Culm saddle. This idea is supported only by one person, namely Williams, for quite a short time.
7. „At locality x in Devonshire, the geological sequence is unbroken, that is, there is no gap“. During the whole debate, such sentences are highly controversial because of being easily classified as local peculiarities.
8. „At locality x in Devon, there is a gap in the geological sequence ranging from period y to z“.

In order to save sentences, there is also merging of sentences. This procedure is problematic inasmuch as it diminishes the historical adequateness of a person's position.

Consider the following example: "Concerning the Non-Culm strata, the amount of fossils under study is sufficiently large." This sentence replaces the two following ones:

1. "Concerning some Non-Culm strata in north Devon, the amount of fossils under study is sufficiently large."
2. "Concerning some Non-Culm strata in South Devon, namely the Great Limestones, the amount of fossils under study is sufficiently large."

For a certain time interval, it seems historically quite adequate to suppose that De La Beche accepts the first and rejects the second of these sentences.

B.2. Interpretative Schemes and Dating Hypotheses

As stated above, the reconstruction of the great Devonian controversy relies heavily on (Rudwick, 1988). There, the debate is analyzed in terms of *interpretative schemes* which are diagrammatically summarized in Fig. B.2.

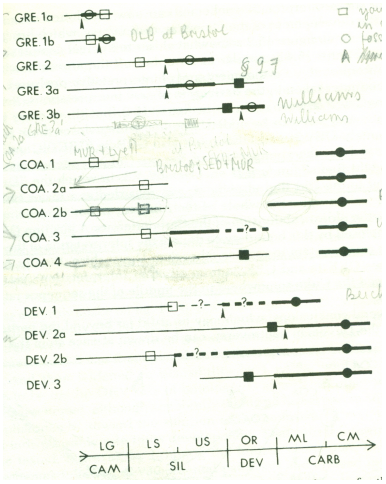

Figure B.1.: Interpretative Schemes (Rudwick, 1988, S.407). Diagrammatic survey of schemes interpreting the older strata of Devonshire. Every horizontal line represents a geological sequence of strata. The geological timescale is shown at the bottom. Thick and thin indicate the Culm strata and the Non-Culm strata, respectively.

For analytical reasons, I analyze the great Devonian controversy in terms of *dating hypotheses* and *bodies of evidence*. Note that, for most dating hypotheses, there is a corresponding *interpretative scheme* delivering the same information. Exceptions are those dating hypotheses which do not determine a certain time order of all the older strata of Devonshire. So, for example, in terms of dating hypotheses, there is no difference between the interpretative schemes $GRE.1a$ and $GRE.1b$. However, for some time steps and participants, the body of evidence contains a sentence

stating a certain time order. In these cases, based on the dating hypothesis as well as the body of evidence, a certain interpretative scheme is singled out. For example, at $S0$, De la Beche considers all the older strata as Cambrian in age. So, his dating hypothesis does not differentiate between $GRE.1a$ and $GRE.1b$. However, De la Beche presupposes a certain time order of all the older strata of Devonshire, namely that some Non-Culm strata in north Devon are the youngest. So, taking his evidential beliefs into account, it is clear that De la Beche interprets the older strata of Devonshire in accordance with $GRE.1a$ and not $GRE.1b$.

For six main participants, Fig. B.2 and Fig. B.3 show time slicing with respect to the interpretative schemes and dating hypotheses, respectively. In both cases, $S0, .., S8$ are time steps of the debate. The six main participants are denoted by DLB, MUR, LYE, SED, PHI and AUS.

In accordance with (Rudwick (1988), especially p. 407 as well as p. 412/3), Fig. B.2 shows interpretation schemes sorted by time and persons. In accordance with (Rudwick, 1988, p. 407) and (Rudwick, 1988, p. 412), Fig. B.3 shows dating hypotheses, also sorted by time and persons.[1] All the older strata which have to be dated are partitioned into three parts, namely the main part of the Culm strata (MC), its black limestone (BCL) and the Non-Culm strata (NC). For each part, there are the same 5 possible geological ages, namely Coal Measures (CM), Mountain Limestone (ML), Old Red Sandstone (ORS), Silurian (SIL) and Cambrian (CAM). Hence, a complete dating hypothesis consists of 15 so-called atomic dating hypotheses (compare sec. 3.1.2).

[1] For analytical reasons, at $S3$, only one dating hypothesis is attributed to Murchison, not two as Rudwick (1988) would suggest.

Time	DLB	MUR	LYE	PHI	SED	AUS
S0	GRE.1a					
S1	GRE.1a	COA.1	COA.1			
S2	GRE.1a	COA.1	COA.1	GRE.1a		
S3	GRE.1b	COA.3, COA.2a	COA.2a	GRE.1b	COA.2a	
S4	GRE.2	COA.2a	COA.2a	COA.2b	COA.2b	
S5	GRE.2	COA.4	COA.2a	COA.2b	COA.2b	
S6	GRE.3a	COA.2b	COA.2b	COA.2b	DEV.2b	DEV.2a
S7	GRE.3a	DEV.3	DEV.3	COA.2b	DEV.2b	DEV.2a
S8	DEV.3	DEV.3	DEV.3	DEV.3	DEV.3	DEV.3

Figure B.2.: Dynamics of interpretation schemes. For six main participants (DLB, MUR, LYE, SED, PHI and AUS), the interpretative scheme is shown at 9 time steps (S0,...,S8).

Figure B.3.: Dynamics of dating hypotheses. For six main participants (DLB, MUR, LYE, SED, PHI and AUS), the dating hypothesis is shown at 9 time steps (S0,...,S8). The strata which have to be dated comprise the main part of the Culm strata (MC), its black limestone (BCL) and the Non-Culm strata (NC). For each part, there are the same 5 possible geological ages, namely Coal Measures (CM), Mountain Limestone (ML), Old Red Sandstone (ORS), Silurian (SIL) and Cambrian (CAM).

B.3. Bodies of Evidence

As already stated in sec. 3.1.3, for every time step and main participant, there is a set of initial beliefs, the so-called body of evidence consisting of empirical and non-empirical statements. In my reconstruction, for every sentence, there is an abbreviation, a so-called sentence title. A body of evidence comprises beliefs which are shared by all participants all of the time, so-called shared evidential beliefs, and others, so-called individual evidential beliefs.

In the following, for every time step, shared evidential beliefs and a participant's individual evidential beliefs are shown. Fig. B.4 and Fig. B.5 show the dynamics of shared evidential beliefs. The other six plots of this subsection show the dynamics of individual evidential beliefs, each corresponding to one of the main participants.

In case of shared evidential beliefs, sentence titles as well as corresponding sentences are displayed. In case of a participant's individual evidential beliefs, only sentence titles are displayed. The corresponding sentences are listed in sec. B.4. A sentence title is colored black, green or red according to whether the corresponding sentence is newly added, removed, or its truth-value assignment has been reversed. Exclamation marks are indicating the rejection of a sentence.

Time	Shared Non−Empirical Statements	Shared Empirical Statements
S0	[Standard Sequence]: Given the strata are undisturbed, Primary strata are overlain by Cambrian strata, Silurian strata, Old Red Sandstone strata, Mountain Limestone strata, Coal Measures strata, New Red Sandstone strata and Oolitic stata, respectively. [Formation of Strata]: Strata are originally formed from sediments that were deposited in flat horizontal sheets − the younger sediments deposited on older ones. [MC as Post−Primary and Pre−NRS]: The main part of the Culm strata is older than Primary strata and younger than New Red Sandstone strata. [BCL as Post−Primary and Pre−NRS]: The black Culm limestone is older than Primary strata and younger than New Red Sandstone strata. [NC as Post−Primary and Pre−NRS]: The Non−Culm strata are older than Primary strata and younger than New Red Sandstone strata. [BCL Older Than MC]: The black Culm limestone is the lowest part of the Culm series. [Conformable Passage − MC and BCL]: The passage between the two parts of the Culm strata is conformable.	[CM Plants in Devon Culm]: Near Bideford (North Devon), there are Coal Measures fossil plants in the main part of the Culm strata.
S1		[Scottish ORS − Rocks]: In Scotland, strata, being Old Red Sandstone in age, support a distinct rock type, namely red sandstone.
S2	[Pembrokeshire − CM in Age]: Near Pembrokeshire (southwest Wales), strata, looking like strata older than Old Red Sandstone, are Coal Measures in age.	[LV in Fauna and Flora −Today]: Today, there are local variations in fauna and flora. [Pembrokeshire − Appearance]: Near Pembrokeshire (southwest Wales), some strata are older than Old Red Sandstone in appearance.

Figure B.4.: Shared evidential beliefs - Part I.

Time	Shared Non−Empirical Statements	Shared Empirical Statements
S3	[Culm Trough as Best Explanation]: Assuming that, originally, younger strata rest upon older ones, the existence of a Culm trough is the best explanation of the Culm strata being juxtaposed solely with older strata.	[Culm trough in Central Devon]: In central Devon, there is a Culm trough−formed by a band of black limestone outcropping south of Barnstaple and north of Dartmoor.
S4	[Yorkshire ML − FA − No LV]: In Yorkshire, the fossils assemblage of strata, which are Mountain Limestone in age, is no local variation, that is (i) the region under study as well as (ii) the amount of fossils is sufficiently large. [BCL − FA − No LV]: The fossils assemblage the black Culm limestone is no local variation, that is (i) the region under study as well as (ii) the amount of fossils is sufficiently large. [Scottland − Body of Evidence − Fossils]: Concerning Scottish Old Red Sandstone strata, the amount of fossils under study is sufficiently large	[ML and Pennines]: All over the Pennines, considering strata which are Mountain Limestone in age, there is a huge lateral variation in rock type.
S5		[No Holoptychius in North Devon]: In north Devon, the Non−Culm strata do not support peculiar fish fossils, like for example the Holoptychius.
S6		
S7	[MC− FA − No LV]: The fossils assemblage of the main part of the Culm strata is no local variation, that is (i) the region under study as well as (ii) the amount of fossils is sufficiently large	
S8	[Russian ORS − Body of Evidence]: Concerning Russian Old Red Sandstone strata, the amount of fossils and rock specimens as well as the region under study is sufficiently large [NC and Scottish Strata − Body of Evidence]: Considering the Non−Culm strata as well as Scottish strata, the amount of fossils and rock specimens as well as the region under study are sufficiently large.	[Valday Hills]: In the Valday hills (Russia), some strata, sandwiched between undisturbed Silurian and Mountain Limestone strata, support limestones with fossils intermediate in character between those of Silurian strata and Mountain Limestone strata as well as sandstones with peculiar Old Red Sandstone fish fossils.

Figure B.5.: Shared evidential beliefs - Part II.

DLB Time	Non–Empirical Statements	Empirical Statements
S0	[Carb Fossils in Pre–ORS strata] [Characteristic Rock Type Principle]	[Unbroken Sequence of Devon Strata] [Passing Devon Northwards] [MC CRT – CAM] [BCL CRT – CAM] [NC CRT – CAM]
S1	! [Characteristic Fossil Principle] ! [Characteristic CM Fossils in Main Culm] ! [Characteristic Fossil Assemblage Principle] ! [CFA – ORS] ! [CFA – ORS – II]	[Scottish ORS – Fossils]
S2	[LV in Fauna and Flora] [LV in Sedimentation] [Characteristic Rock Type Principle] [Characteristic Fossil Principle] [Characteristic CM Fossils in Main Culm] [Characteristic Fossil Assemblage Principle] [CFA – ORS] [CFA – ORS – II]	[Northwest France] [MC CRT – CAM] [BCL CRT – CAM] [NC CRT – CAM]
S3	[Main Culm Youngest Devonian Strata]	! [Passing Devon Northwards]
S4	! [Lyellian Principle – V2] [Characteristic Fossil Assemblage Principle – V2] ! [Scottland – Body of Evidence – Region]	! [BCL – FA]
S5	! [Non–Culm – Body of Evidence – Fossils]	[Carboniferous Plants in North Devon]
S6	[CFA – ORS – II – V2] [South Devon LSTs – FA – 2] [Devon Strata – Temporal Order – 3] [Non–Culm – Body of Evidence – Region] [Non–Culm – Body of Evidence – Fossils] [Scottland – Body of Evidence – Region]	[South Devonian Fossil Fauna]
S7	[Devon Strata – Temporal Order – 3]	! [Non–Culm Fossil Mixture – Devon] [Sequence of Strata – Tor Bay and Newton Abott] ! [MC – FA]
S8	[Main Culm Youngest Devonian Strata] [Non–Culm – Body of Evidence – Region] [CFA – ORS – III – V2] [CFA – ORS – II – V2] [South Devon LSTs – FA – 2]	[Philipps' Collection] ! [Sequence of Strata – Tor Bay and Newton Abott] [Non–Culm Fossil Mixture – Devon] [MC – FA] [BCL – FA]

Figure B.6.: Dynamics of De la Beche's individual evidential beliefs. Displayed are sentence titles. A listing of sentences and corresponding titles is given in sec. B.4. A sentence title is colored black, green or red according to whether the corresponding sentence is newly added, removed, or its truth-value assignment has been reversed. Exclamation marks are indicating the rejection of a sentence.

MUR		
Time	Non–Empirical Statements	Empirical Statements
S1	! [Carb Fossils in Pre–ORS strata] ! [Characteristic Rock Type Principle] [Characteristic Fossil Principle] [Characteristic CM Fossils in Main Culm] ! [Characteristic Fossil Assemblage Principle] [CFA – ORS] [CF – ORS]	[Scottish ORS – Fossils] ! [MC CRT – CAM] ! [BCL CRT – CAM] ! [NC CRT – CAM]
S2	! [LV in Fauna and Flora] [LV in Sedimentation] [Characteristic Rock Type Principle]	! [Northwest France] [MC CRT – CAM] [BCL CRT – CAM] [NC CRT – CAM]
S3	[Main Culm Youngest Devonian Strata]	! [Passing Devon Northwards] [SIL Fossils in North Devon]
S4	! [Lyellian Principle – V2] ! [Characteristic Fossil Assemblage Principle – V2] [Characteristic Fossil Asseblage – Some Time] [Scottland – Body of Evidence – Region] [Characteristic Fossil Assemblage Principle]	
S5	[Non–Culm – Body of Evidence – Fossils] [Non–Culm – Body of Evidence – Region] [Characteristic Fossil – Some Time] ! [Characteristic Fossil Principle] [CF – ORS]	[Carboniferous Plants in North Devon]
S6	[Characteristic Fossil Principle] [CF – ORS] [Non–Culm – Body of Evidence – Region]	[BCL – FA] ! [South Devonian Fossil Fauna] ! [Carboniferous Plants in North Devon]
S7	[NC – FA – 2] [Lyellian Principle – V2] [CF – ORS] [Characteristic Fossil Principle] [CFA – ORS] [Characteristic Fossil Assemblage Principle – V2] [Characteristic Fossil Asseblage – Some Time]	[Non–Culm Fossil Mixture – Devon] [MC – FA] ! [Sequence of Strata – Tor Bay and Newton Abott] [South Devonian Fossil Fauna] [Carboniferous Plants in North Devon] ! [SIL Fossils in North Devon]
S8	[Characteristic Fossil Asseblage – Some Time] [CFA – ORS – III – V2] ! [Scottland – Body of Evidence – Region] ! [Non–Culm – Body of Evidence – Region]	[Philipps' Collection] [Non–Culm Fossil Mixture – Devon] [MC – FA] [BCL – FA]

Figure B.7.: Dynamics of Murchison's individual evidential beliefs. Displayed are sentence titles. A listing of sentences and corresponding titles is given in sec. B.4. A sentence title is colored black, green or red according to whether the corresponding sentence is newly added, removed, or its truth-value assignment has been reversed. Exclamation marks are indicating the rejection of a sentence.

LYE		
Time	Non–Empirical Statements	Empirical Statements
S1	! [Characteristic Rock Type Principle] ! [Characteristic Fossil Principle] ! [Characteristic CM Fossils in Main Culm] [Lyellian Principle] [Carb and Pre–ORS – Sufficiently Different] ! [CF – ORS]	! [Scottish ORS – Fossils] ! [MC CRT – CAM] ! [BCL CRT – CAM] ! [NC CRT – CAM]
S2	! [LV in Fauna and Flora] [LV in Sedimentation] ! [Characteristic Fossil – Some Time] [CF – ORS] [Characteristic Rock Type Principle] [Characteristic Fossil Principle] [Characteristic CM Fossils in Main Culm]	! [Northwest France] [MC CRT – CAM] [BCL CRT – CAM] [NC CRT – CAM]
S3	[Main Culm Youngest Devonian Strata]	! [Passing Devon Northwards]
S4		! [BCL – FA]
S5	[Non–Culm – Body of Evidence – Fossils] [Non–Culm – Body of Evidence – Region]	! [Carboniferous Plants in North Devon]
S6		! [South Devonian Fossil Fauna] [BCL – FA]
S7	[NC – FA – 2]	[Non–Culm Fossil Mixture – Devon] [MC – FA] ! [Sequence of Strata – Tor Bay and Newton Abott] [South Devonian Fossil Fauna] [Carboniferous Plants in North Devon]
S8	! [Carb Fossils in Pre–ORS strata] [Lyellian Principle – V2] [Characteristic Fossil Asseblage – Some Time] [CFA – ORS – III – V2] ! [Non–Culm – Body of Evidence – Region] [Lyellian Principle]	[Philipps' Collection] [Scottish ORS – Fossils] [Non–Culm Fossil Mixture – Devon] [MC – FA] [BCL – FA]

Figure B.8.: Dynamics of Lyell's individual evidential beliefs.. Displayed are sentence titles. A listing of sentences and corresponding titles is given in sec. B.4. A sentence title is colored black, green or red according to whether the corresponding sentence is newly added, removed, or its truth-value assignment has been reversed. Exclamation marks are indicating the rejection of a sentence.

PHI		
Time	Non–Empirical Statements	Empirical Statements
S2	[Carb Fossils in Pre–ORS strata] [LV in Fauna and Flora] [LV in Sedimentation]	[Scottish ORS – Fossils] [Northwest France]
S3	[Main Culm Youngest Devonian Strata]	! [Passing Devon Northwards]
S4	[Characteristic Fossil Assemblage Principle – V2] ! [Lyellian Principle – V2] ! [Scottland – Body of Evidence – Region]	[BCL – FA]
S5	! [Non–Culm – Body of Evidence – Fossils]	[Carboniferous Plants in North Devon]
S6		! [South Devonian Fossil Fauna]
S7		! [Non–Culm Fossil Mixture – Devon] [MC – FA] ! [Sequence of Strata – Tor Bay and Newton Abott]
S8	[Lyellian Principle – V2] [Non–Culm – Body of Evidence – Fossils] [CFA – ORS – III – V2] ! [Non–Culm – Body of Evidence – Region]	[Philipps' Collection] [MC – FA] [BCL – FA] [Non–Culm Fossil Mixture – Devon] [South Devonian Fossil Fauna]

Figure B.9.: Dynamics of Phillips's individual evidential beliefs. Displayed are sentence titles. A listing of sentences and corresponding titles is given in sec. B.4. A sentence title is colored black, green or red according to whether the corresponding sentence is newly added, removed, or its truth-value assignment has been reversed. Exclamation marks are indicating the rejection of a sentence.

SED Time	Non–Empirical Statements	Empirical Statements
S3	[Carb Fossils in Pre–ORS strata] [LV in Fauna and Flora] [LV in Sedimentation] [Main Culm Youngest Devonian Strata]	! [SIL Fossils in North Devon] [Scottish ORS – Fossils] [Northwest France] ! [Passing Devon Northwards]
S4	! [Characteristic Fossil Assemblage Principle – V2] ! [Lyellian Principle – V2] [Scottland – Body of Evidence – Region] [CFA – ORS – V2]	[BCL – FA]
S5	[Non–Culm – Body of Evidence – Fossils] [Non–Culm – Body of Evidence – Region]	[Carboniferous Plants in North Devon]
S6		! [South Devonian Fossil Fauna]
S7		! [Non–Culm Fossil Mixture – Devon] [MC – FA] ! [Sequence of Strata – Tor Bay and Newton Abott]
S8	[Characteristic Fossil Assemblage Principle – V2] ! [Non–Culm – Body of Evidence – Region] ! [Scottland – Body of Evidence – Region] [CFA – ORS – III – V2] [CFA – ORS – V2]	[Philipps' Collection] [MC – FA] [BCL – FA] [Non–Culm Fossil Mixture – Devon] [South Devonian Fossil Fauna]

Figure B.10.: Dynamics of Sedgwick's individual evidential beliefs. Displayed are sentence titles. A listing of sentences and corresponding titles is given in sec. B.4. A sentence title is colored black, green or red according to whether the corresponding sentence is newly added, removed, or its truth-value assignment has been reversed. Exclamation marks are indicating the rejection of a sentence.

AUS		
Time	Non–Empirical Statements	Empirical Statements
S6	[Carb Fossils in Pre–ORS strata] [LV in Fauna and Flora] [LV in Sedimentation] [Main Culm Youngest Devonian Strata] [Characteristic Fossil Assemblage Principle – V2] ! [Lyellian Principle – V2] ! [Scottland – Body of Evidence – Region] [CFA – ORS – V2] [Non–Culm – Body of Evidence – Fossils] [Non–Culm – Body of Evidence – Region] [South Devon LSTs – FA – 2]	[South Devonian Fossil Fauna] [Carboniferous Plants in North Devon] [BCL – FA] [Scottish ORS – Fossils] [Northwest France] ! [Passing Devon Northwards] [Unbroken Sequence of Devon Strata]
S7		! [Non–Culm Fossil Mixture – Devon] [MC – FA] ! [Sequence of Strata – Tor Bay and Newton Abott]
S8	! [Non–Culm – Body of Evidence – Region] [CFA – ORS – III – V2] [CFA – ORS – V2]	[Philipps' Collection] [MC – FA] [BCL – FA] [Non–Culm Fossil Mixture – Devon] [Unbroken Sequence of Devon Strata] [South Devon LSTs – FA – 2]

Figure B.11.: Dynamics of Austen's individual evidential beliefs. Displayed are sentence titles. A listing of sentences and corresponding titles is given in sec. B.4. A sentence title is colored black, green or red according to whether the corresponding sentence is newly added, removed, or its truth-value assignment has been reversed. Exclamation marks are indicating the rejection of a sentence.

B.4. Sentences

Here, a sentence is shown together with its title. However, this is not done for all sentences of the reconstruction, but rather for all sentences figuring in the bodies of evidence, compare sec. B.3.

Time	Non−Empirical Statements	Empirical Statements
S0	[Carb Fossils in Pre−ORS strata]: In strata older than Old Red Sandstone, there are Carboniferous fossils. [Characteristic Rock Type Principle]: There is a bijection between the age of some strata and its characteristic rock types.	[Unbroken Sequence of Devon Strata]: In Devon, all the strata − older than the New Red Sandstone − form an unbroken sequence of strata. [Passing Devon Northwards]: Passing Devon from south to north, the strata become younger, except the strata north of Exmoor which are of the same age as the strata south of Exmoor [MC CRT − CAM]: The main part of the Culm strata are of characteristic Cambrian rock type. [BCL CRT − CAM]: The black Culm limestone is of characteristic Cambrian rock type. [NC CRT − CAM]: The Non−Culm strata are of characteristic Cambrian rock type.
S1	[Characteristic Fossil Principle]: There is a bijection between the age of some strata and its characteristic fossils. [Characteristic Fossil Assemblage Principle]: There is a bijection between the age of some strata and its characteristic fossils assemblages. [Lyellian Principle]: For two strata A and B it holds: (i) The more A and B are similar in their fossils assemblages, the more they are similar in age, (ii) the more A and B are similar in age, the more they are similar in their fossils assemblages and (iii) if A and B are sufficiently unsimilar in age, then they have no species in common. [Carb and Pre−ORS − Sufficiently Different]: Carboniferous strata and strata older than Old Red Sandstone are sufficiently different. [CFA − ORS]: The characteristic fossils assemblage of Old Red Sandstone strata supports only a few fossils, most of them peculiar fish fossils not known from other strata. [CFA − ORS − II]: The characteristic fossils assemblage of Old Red Sandstone strata is intermediate in character between those of Silurian strata and Mountain Limestone strata [CF − ORS]: Some peculiar fish fossils are characteristic of the Old Red Sandstone period. [Characteristic CM Fossils in Main Culm]: The main part of the Culm strata supports fossil plants which are characteristic of the Coal Measures era.	[Scottish ORS − Fossils]: Scottish strata, which are Old Red Sandstone in age, support only a few peculiar fish fossils not known from other strata (e.g. Holoptychius), lacking in any ordinary marine fossils (such as mollusks, brachiopods and corals).
S2	[LV in Fauna and Flora]: For every period in the history of the earth, there are local variations in fauna and flora. [LV in Sedimentation]: For every period in the history of the earth, there are local variations in sedimentation.	[Northwest France]: In northwest France, there are some strata older than Old Red Sandstone supporting fossil plants which are Coal Measures in age.
S3	[Main Culm Youngest Devonian Strata]: In Devon, among all the strata older than New Red Sandstone, the main part of the Culm strata are the youngest.	[SIL Fossils in North Devon]: In north Devon, near Barnstaple, there are characteristic Silurian fossils in the Non−Culm strata.

Figure B.12.: Sentences figuring in a body of evidence - Part I.

Time	Non−Empirical Statements	Empirical Statements
S4	[Lyellian Principle − V2]: Given a sufficiently large (i) amount of fossils and (ii) region under study, for two strata A and B, it holds: (ia) The more A and B are similar in their fossils assemblages, the more they are similar in age, (ib) if A and B are sufficiently unsimilar in age, then they have no fossil species in common, and (iia) the more A and B are similar in age, the more they are similar in their fossils assemblages. [Characteristic Fossil Assemblage Principle − V2]: Given a sufficiently large (i) amount of fossils and (ii) region under study, it holds: There is a bijection between the age of some strata and its characteristic fossils assemblages. [Characteristic Fossil Asseblage − Some Time]: For some geological period, there is (at least) one characteristic assemblage of fossil species. [CFA − ORS − V2]: Given a sufficiently large (i) amount of fossils and (ii) region under study, Old Red Sandstone strata encompasses only a few fossils, most of them peculiar fish fossils not known from other strata. [Scottland − Body of Evidence − Region]: Concerning Scottish strata which are Old Red Sandstone in age, the region under study is sufficiently large.	[BCL − FA]: Most of the fossil species supported by the black Culm limestone are known from strata around Yorkshire which are Mountain Limestone in age.
S5	[Non−Culm − Body of Evidence − Fossils]: Concerning the Non−Culm strata, the amount of fossils under study is sufficiently large. [Non−Culm − Body of Evidence − Region]: Concerning the Non−Culm strata, the region under study is sufficiently large. [Characteristic Fossil − Some Time]: In north Devon, near Barnstaple, there are characteristic Silurain fossils in the Non−Culm strata.	[Carboniferous Plants in North Devon]: In north Devon, some Non−Culm strata support Carboniferous plant fossils.
S6	[CFA − ORS − II − V2]: Given a sufficiently large (i) amount of fossils and (ii) region under study, strata, which are Old Red Sanstone in age, support a fossils assemblage intermediate in character between those of strata, which are Silurian and Mountain Limestone in age. [South Devon LSTs − FA − 2]: South Devonian limestones, around Tor Bay and Netwon Abott, support a fossils assemblage more similar to that of Mountain limestone strata than that of Silurian strata. [Devon Strata − Temporal Order − 3]: The main part of the Non−Culm strata are older than the black Culm limestone, the main part of the Culm strata and some South Devonian limestones, respectively.	[South Devonian Fossil Fauna]: Around Tor Bay and Netwon Abott, the limestones support not one single Silurian fossil (and no peculiar fish fossil like for example the Holoptychius), but many Mountain Limestone shells and a great many new species.

Figure B.13.: Sentences figuring in a body of evidence - Part II.

Time	Non−Empirical Statements	Empirical Statements
S7	[NC − FA − 2]: Non−Culm strata support fossils intermediate in character between those of Silurian and Mountain Limestone strata.	[Non−Culm Fossil Mixture − Devon]: In north Devon, the Non−Culm fossils are a mixture of Mountain Limestone shells, Silurian corals, one Old Red Sandstone mollusk and Carboniferous plants; some trilobites similar to Silurian ones and some totally new fossils; in south Devon, few Non−Culm fossils are strictly Mountain Liemstone fossils; many fossils are new. [Sequence of Strata − Tor Bay and Newton Abott]: Around Tor Bay and Newton Abott, (i) strata are undisturbed and (ii) some limestones overlay the main part of the Culm strata, the black Culm limestone and the main part of the Non−Culm strata, respectively [MC − FA]: Most of the fossil species supported by the main part of the Culm strata are known from the Coal measures era.
S8	[CFA − ORS − III − V2]: Given a sufficiently large (i) amount of fossils and (ii) region under study, it holds: Old Red Sandstone strata support a characteristic fossils assemblage intermediate in character between those of Silurian strata and Mountain Limestone strata as well as peculiar fish fossils only known from Old Red Sandstone strata.	[Philipps' Collection]: Most of the fossil species supported by the main part of the Culm and the black Culm limestone a are known from strata which are Coal Measures and Mountain Limestone in age, respectively; Non−Culm strata support a fossils assemblage intermediate in character between those of Silurian strata and Mountain Limestone strata.

Figure B.14.: Sentences figuring in a body of evidence - Part III.

C. Polarization Dynamics

C.1. Theoretical Distance and Unsimilarity

In this thesis, polarization is assessed in terms of groups. Groups are defined endogenously using different similarity thresholds. Similarity (SIM) is considered as a function with two arguments ranging from 0 to 1, see definition 3.2.2. Two dating hypotheses are more similar, if they share more atomic dating hypotheses. To each similarity measure, there is a corresponding unsimilarity measure ($1 - SIM$). Comparing two dating hypotheses, the unsimilarity measure is a normalized Hamming distance. Fig. 3.7 shows similarity dynamics for each of pair of two main participants.

Fig. 3.10 shows clustering results for a similarity threshold of 1.0. Considering dating hypotheses, there are groups and they are the same as those in Fig. 3.2, which is taken from (Rudwick, 1988). Fig. 3.2 shows not only groups, but also the so-called *theoretical distance* between two persons. According to Rudwick (1988), it is unquantifiable. Is there some relation between *theoretical distance* and similarity respectively unsimilarity of dating hypotheses? Comparison of Fig. 3.2 and Fig. 3.7 shows that, for De la Beche and Murchison, there are differences between these two concepts. These differences are listed below:

1. The maximum of *theoretical distance* is reached right at the beginning at time step 1. The unsimilarity reaches its maximal value not at the beginning, but at time step 7.

2. Moving from time step 2 to 3, theoretical distance decreases, but unsimilarity increases.

3. Neglecting the final step, the minimum of *theoretical distance* is reached at time step 5. Neglecting the final step, unsimilarity reaches its minimal value not at time step 5, but at time steps $1, 2$ and 4.

See Fig. B.2 and Fig. B.3 to find the corresponding *interpretative schemes* and dating hypotheses. As *theoretical distance* is not fully captured by the unsimilarity measure, let us take a look at some additional characteristics of *interpretative schemes*, compare sec. B.2.

Some words on differences in sizing a gap in the temporal sequence of all the older strata in Devon. For De la Beche and Murchison, it shows that these differences are biggest right at the beginning. De la Beche denies the existence of a gap and Murchison states a gap comprising 3 geological ages. As time goes by, De la Beche constantly holds on to his belief, while Murchison successively decreases the size of the gap. Moving from time step 2 to 3, these differences decrease. Neglecting the final step, differences are smallest at time step 7, with both, De la Beche and Murchison, agreeing on there being no gap in the sequence.

Some words on the temporal order of all the older strata in Devon. For large parts of the debate, De la Beche places Culm strata in the middle of the sequence, namely from the beginning until time step 3 and from time step 6 to the penultimate time step. Murchison constantly places Culm strata on top of the sequence. Therefore, moving from time step 2 to 3, De la Beche and Murchison no longer differ on that point. However, this agreement already ends with time step 6.

As a result of this short analysis, it seems that *theoretical distance* is assessed comparing (i) the temporal order of all the older strata in Devon as well as (ii) the size of a gap in the sequence of the same strata.

C.2. Additional Similarity Clustering Results

Comparing five different similarity thresholds, namely 0.8, 0.85, 0.9, 0.95 and 1, it shows that similarity clusters depend on the similarity threshold. In the following, results relying on a similarity threshold of 0.8, 0.9 and 0.95 will be presented in some more detail. See sec. 3.2.1 for results relying on a similarity thresholds 0.85 and 1.

First, a short summary of the results for a similarity threshold of 0.95, compare Fig. C.3 and Fig. C.6. Considering dating hypotheses, results are the same as for a similarity threshold of 1. Considering bodies of evidence, results are not the same as for a similarity threshold of 1. There are groups, albeit not many and only small ones. First, there is only one group, namely De la Beche and Phillips. Entering the

stage, Sedgwick joins this group, but only for a short while, namely two adjacent time steps. At time step 6*b*, De la Beche and Phillips split, that is, there is no group left. Moving one step forward, there is a new group, namely De la Beche and Austen. However, they split already after one time step, leaving no group at all until the final time step. Finally, there are three groups, namely (*i*) Murchison and Lyell, (*ii*) De la Beche, Sedgwick and Austen, and (*iii*) Phillips, Sedgwick and Austen. Being not similar enough, De la Beche and Phillips cannot form a group together with Sedgwick and Austen. Similarity clusters change with changes in the body of shared background beliefs. As already mentioned, moving from time step 6*b* to 7*a*, a new group pops up, namely De la Beche and Austen.

Second, a short summary of the results for a similarity threshold of 0.9, compare Fig. C.2 and Fig. C.5. Considering dating hypotheses, first, results are the same as for similarity thresholds of 1 and 0.95. From time step 5*b* till the end, there are some differences. At 5*b*, there is a new group consisting of Murchison and Lyell. Here, Murchison and Lyell differ only in accepting a single atomic dating hypothesis, namely that some Non-Culm strata are Old Red Sandstone in age. At time steps 6*b*, Austen joins the group of Murchison, Lyell and Phillips only differing in accepting the same atomic dating hypothesis. However, only after two steps, this groups breaks into two, namely (*i*) Murchison and Lyell and (*ii*) Phillips and Austen. Considering bodies of evidence, first, results are the same as for similarity thresholds of 1 and 0.95. From time step 4*b* till the end, there are some difference. At 4*b*, there is a new group consisting of Phillips and Sedgwick. At time step 6*b*, there are even two new groups popping up, namely Austen and De la Beche as well as Austen and Phillips. However, these new groups do not last very long, indeed only one step and two steps, respectively. At time steps 7*b* and 8*a*, instead of no groups at all, there are several groups. At both time steps, there is a group consisting of Murchison and Lyell as well as another group consisting of Austen and De la Beche. Only at 8*a*, Sedgwick and Phillips are similar enough to form a group together with Austen. Finally, persons are divided into two groups, that is, Murchison and Lyell oppose all the others. Similarity clusters change with changes in the body of shared background beliefs. As already mentioned, moving from time step 7*b* to 8*a*, Sedgwick and Phillips get more similar.

Third, a short summary of the results for a similarity threshold of 0.85, compare Fig. 3.9 and Fig. 3.11. Considering dating hypotheses, first, results are the same as for a similarity threshold of 0.9. From time step 4*b*till the end, there are some

differences. At 4*b*, two small groups merge, namely (*i*) Murchison and Lyell and (*ii*) Phillips and Sedgwick. However, with the next individual belief change, Murchison splits. At time step 6*b*, Sedgwick is now sufficiently similar to Murchison, Lyell and Phillips. Therefore, there is additionally a new group consisting of all the four of them. At time step 7*b*, the group consisting of Murchison and Lyell is joined by Austen. Additionally, there is a new group consisting of Sedgwick and Phillips. Considering bodies of evidence, first, results are the same as for a similarity threshold of 0.9. From time step 4*a* till the end, there are some differences. At 4*a*, there is a new group consisting of Lyell and Murchison. At time step 4*b*, two groups merge, which do not split until time step 6*b*, namely (*i*) De la Beche and Phillips and (*ii*) Phillips and Sedgwick. At time step 6*b*, there are some more differences. Using a similarity threshold of 0.85, De la Beche and Austen are joined by Phillips, who additionally forms a new group with Sedgwick and Austen. At time step 7*a*, with Sedgwick being sufficiently similar to De la Beche, these two groups merge, but not for long. Already at the next step, this group breaks into two, namely (*i*) De la Beche and Austen and (*ii*) Austen, Phillips and Sedgwick. From time step 6*b* until the end and additionally at time step 4*a*, Murchison and Lyell are sufficiently similar. Probably, the most remarkable difference between using a similarity threshold of 0.9 and 0.85 can be observed at the final step. Only using a similarity threshold of 0.85, all pairs of persons are sufficiently similar, that is, there is a group uniting all persons.

Fourth, a short summary of the results for a similarity threshold of 0.8, compare Fig. C.1 and Fig. C.4. Considering dating hypotheses, first, results are the same as for a similarity threshold of 0.85. At time steps 5*b* and 6*a*, there is some difference, namely Murchison not splitting being still sufficiently similar to Phillips and Sedgwick. At time steps 6*b*and 7*a*, results differ in Sedgwick being sufficiently similar with Austen. Only using a similarity threshold of 0.8, there is a group consisting of all persons but De la Beche. At time step 7*b*and 8*a* results differ once more in (*i*) Phillips being part of the group consisting of Murchison, Lyell and Austen and (*ii*) Sedgwick being part of the group consisting of Phillips and Austen. Considering bodies of evidence, only at the very beginning, results are the same as for a similarity threshold of 0.85. At time steps 2*b*, 3*a* and 3*b*, there is some difference, namely a new group. At 2*b*, the new group consists of Murchison and Lyell. At 3*a* and 3*b*, the new group consists of Murchison, Lyell and Phillips. At 3*b*, results differ additionally in Lyell being clustered together with De la Beche, Phillips and Sedgwick.

At time step $4a$, only using a similarity threshold of 0.8, there is a group uniting all persons. From $4b$, using a similarity threshold of 0.8, there are additionally several small groups, namely (*i*) Murchison and Lyell, (*ii*) Murchison and Sedgwick, and (*iii*) De la Beche and Lyell. The latter group only lasts one individual belief change. At time step $6b$, results differ in Sedgwick being sufficiently similar to De la Beche respectively Murchison and Lyell. The first part is also true for time step $7a$. At $7b$ and $8a$, there are also differences. Only using a similarity threshold of 0.8, Murchison and Lyell are clustered with Austen respectively Sedgwick. The same is true for De la Beche with regard to Phillips, Sedgwick and Austen.

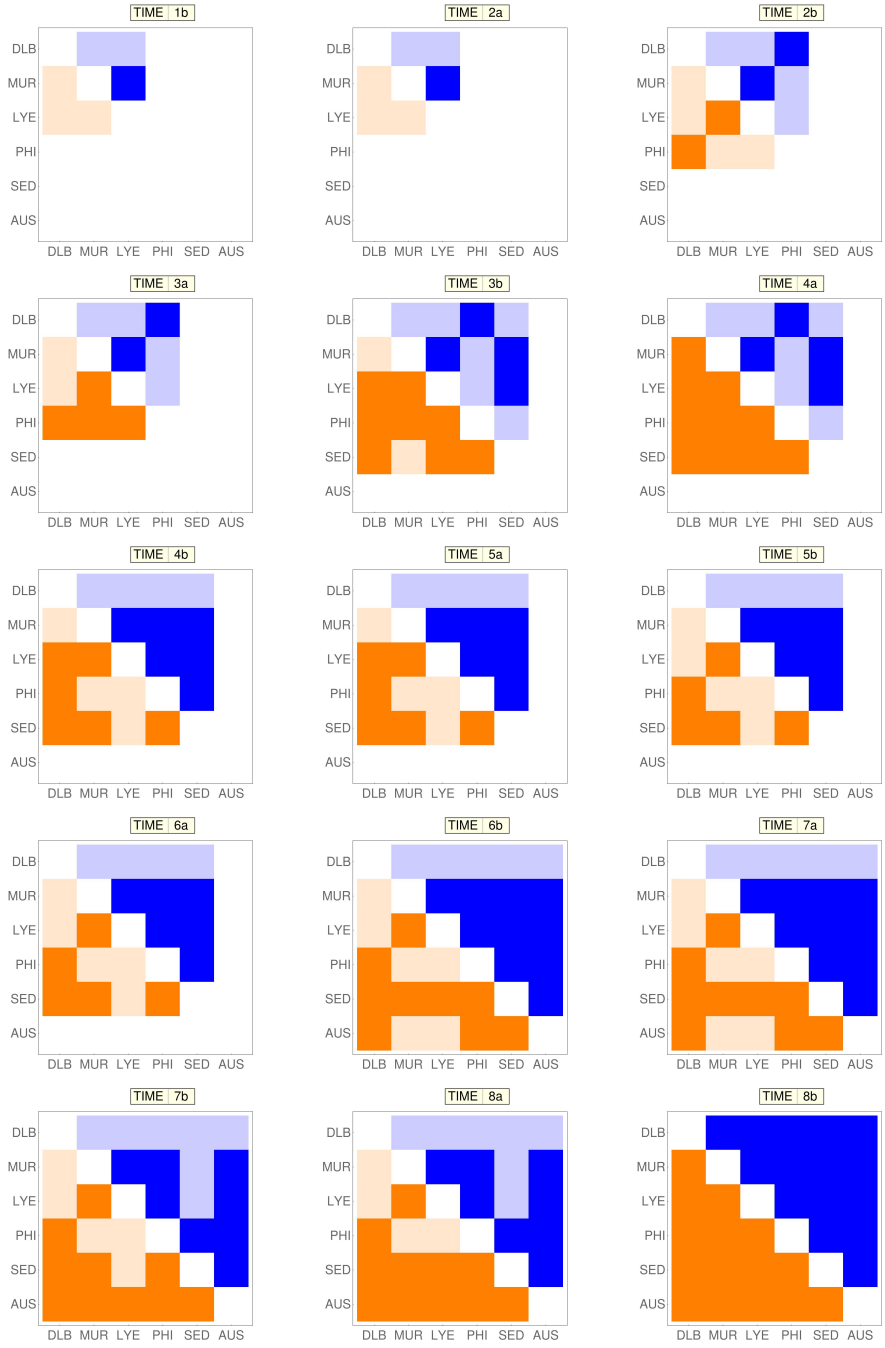

Figure C.1.: A set of plots is shown with every plot corresponding to a certain time step. Every single plot consists of an upper and a lower triangle showing similarity between two dating hypotheses and bodies of evidence, respectively. These hypotheses and bodies of evidence belong to De la Beche (DLB), Murchison (MUR), Lyell (LYE), Sedgwick (SED) and Austen (AUS). Darker and lighter shading represent a similarity of greater or equal 0.8 and less than 0.8, respectively.

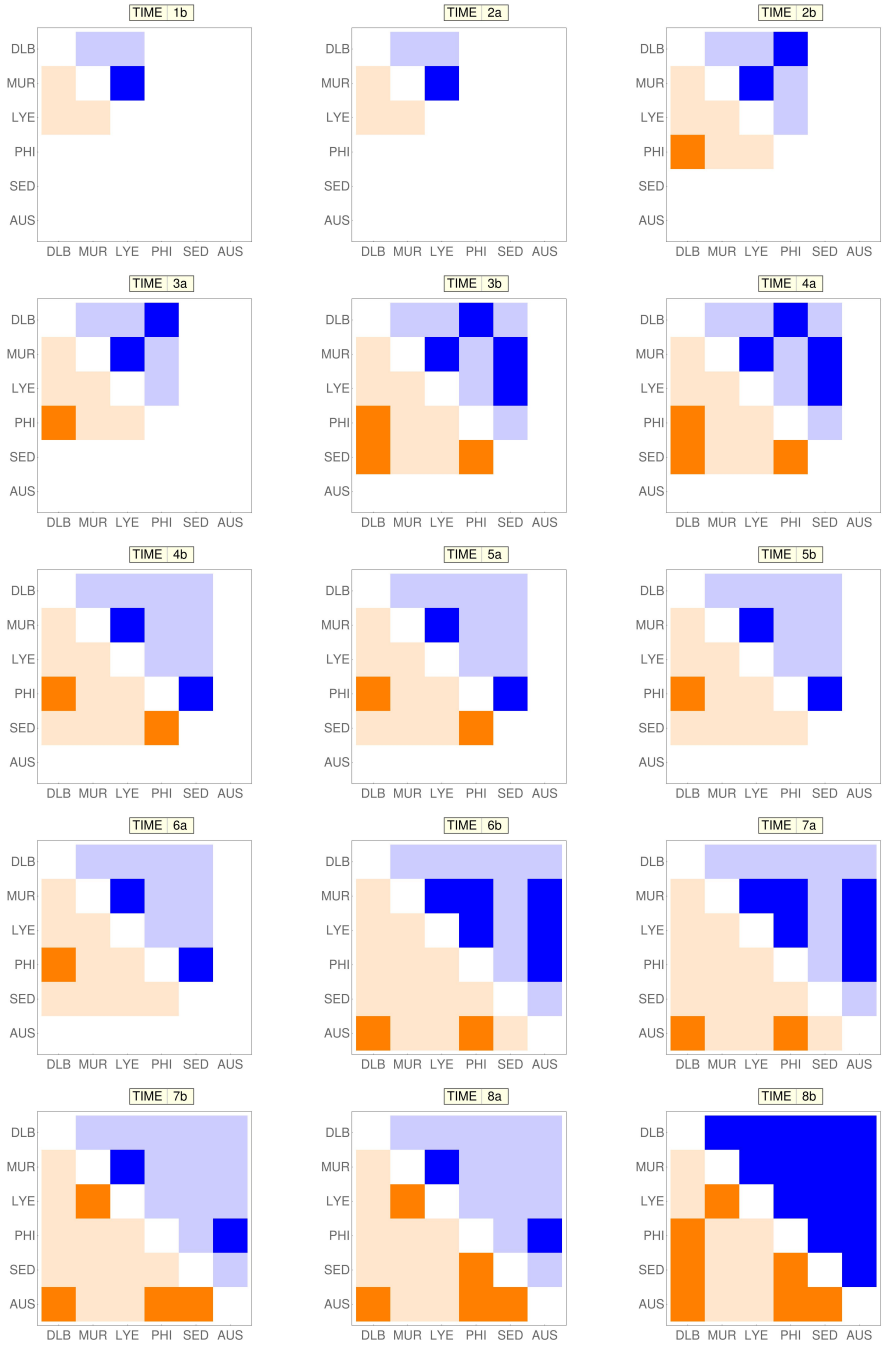

Figure C.2.: A set of plots is shown with every plot corresponding to a certain time step. Every single plot consists of an upper and a lower triangle showing similarity between two dating hypotheses and bodies of evidence, respectively. These hypotheses and bodies of evidence belong to De la Beche (DLB), Murchison (MUR), Lyell (LYE), Sedgwick (SED) and Austen (AUS). Darker and lighter shading represent a similarity of greater or equal 0.9 and less than 0.9, respectively.

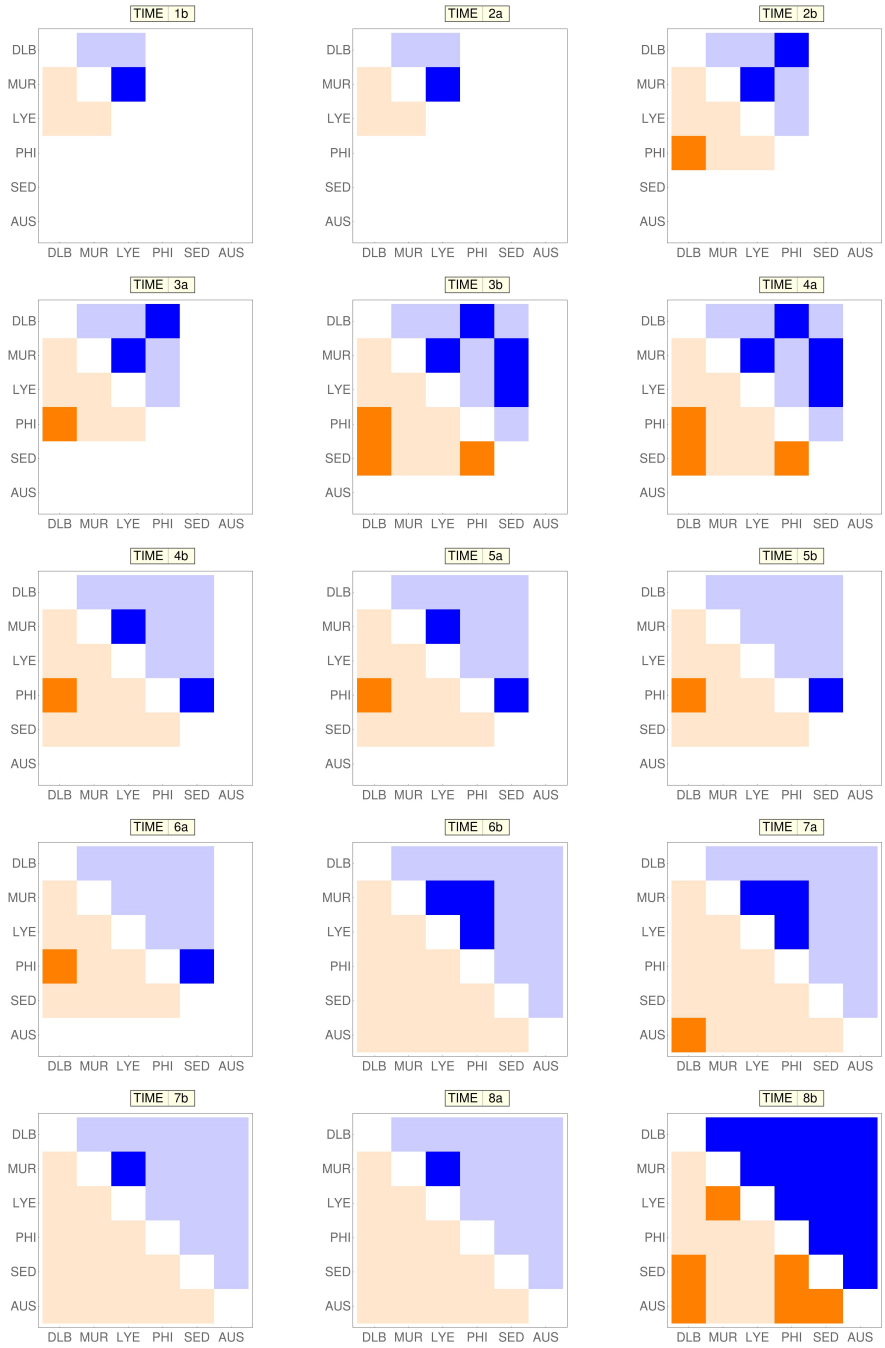

Figure C.3.: A set of plots is shown with every plot corresponding to a certain time step. Every single plot consists of an upper and a lower triangle showing similarity between two dating hypotheses and bodies of evidence, respectively. These hypotheses and bodies of evidence belong to De la Beche (DLB), Murchison (MUR), Lyell (LYE), Sedgwick (SED) and Austen (AUS). Darker and lighter shading represent a similarity of greater or equal 0.95 and less than 0.95, respectively.

SIM≥0.8			
Time	Persons	H–Groups	E–Groups
S1b	DLB, MUR, LYE	(MUR, LYE)	–
S2a	"	"	"
S2b	+ PHI	(MUR, LYE), (DLB, PHI)	(MUR, LYE), (DLB, PHI)
S3a	"	"	(MUR, LYE, PHI), (DLB, PHI)
S3b	+ SED	(MUR, LYE, SED), (DLB, PHI)	(MUR, LYE, PHI), (DLB, LYE, PHI, SED)
S4a	"	"	(DLB, MUR, LYE, PHI, SED)
S4b	"	(MUR, LYE, PHI, SED)	(MUR, LYE), (MUR, SED), (DLB, LYE), (DLB, PHI, SED)
S5a	"	"	"
S5b	"	"	(MUR, LYE), (MUR, SED), (DLB, PHI, SED)
S6a	"	"	"
S6b	+ AUS	(MUR, LYE, PHI, SED, AUS)	(MUR, LYE, SED), (DLB, PHI, SED, AUS)
S7a	"	"	"
S7b	"	(MUR, LYE, PHI, AUS), (PHI, SED, AUS)	(MUR, LYE, AUS), (DLB, PHI, SED, AUS)
S8a	"	"	(MUR, LYE, SED, AUS), (DLB, PHI, SED, AUS)
S8b	"	(DLB, MUR, LYE, PHI, SED, AUS)	(DLB, MUR, LYE, PHI, SED, AUS)

Figure C.4.: Clustering persons according to a similarity threshold of 0.8. Clustering is performed based on similarities between dating hypotheses (H-groups) and bodies of evidence (E-groups).

SIM≥0.9			
Time	Persons	H–Groups	E–Groups
S1b	DLB, MUR, LYE	(MUR, LYE)	–
S2a	"	"	"
S2b	+ PHI	(MUR, LYE), (DLB, PHI)	(DLB, PHI)
S3a	"	"	"
S3b	+ SED	(MUR, LYE, SED), (DLB, PHI)	(DLB, PHI, SED)
S4a	"	"	"
S4b	"	(MUR, LYE), (PHI, SED)	(DLB, PHI), (PHI, SED)
S5a	"	"	"
S5b	"	"	(DLB, PHI)
S6a	"	"	"
S6b	+ AUS	(MUR, LYE, PHI, AUS)	(DLB, AUS), (PHI, AUS)
S7a	"	"	"
S7b	"	(MUR, LYE), (PHI, AUS)	(MUR, LYE), (DLB, AUS), (PHI, AUS), (SED, AUS)
S8a	"	"	(MUR, LYE), (DLB, AUS), (PHI, SED, AUS)
S8b	"	(DLB, MUR, LYE, PHI, SED, AUS)	(MUR, LYE), (DLB, PHI, SED, AUS)

Figure C.5.: Clustering persons according to a similarity threshold of 0.9. Clustering is performed based on similarities between dating hypotheses (H-groups) and bodies of evidence (E-groups)

SIM≥0.95			
Time	Persons	H–Groups	E–Groups
S1b	DLB, MUR, LYE	(MUR, LYE)	–
S2a	"	"	"
S2b	+ PHI	(MUR, LYE), (DLB, PHI)	(DLB, PHI)
S3a	"	"	"
S3b	+ SED	(MUR, LYE, SED), (DLB, PHI)	(DLB, PHI, SED)
S4a	"	"	"
S4b	"	(MUR, LYE), (PHI, SED)	(DLB, PHI)
S5a	"	"	"
S5b	"	(PHI, SED)	"
S6a	"	"	"
S6b	+ AUS	(MUR, LYE, PHI)	–
S7a	"	"	(DLB, AUS)
S7b	"	(MUR, LYE)	–
S8a	"	"	"
S8b	"	(DLB, MUR, LYE, PHI, SED, AUS)	(MUR, LYE), (DLB, SED), (DLB, AUS), (PHI, SED, AUS)

Figure C.6.: Clustering persons according to a similarity threshold of 0.95. Clustering is performed based on similarities between dating hypotheses (H-groups) and bodies of evidence (E-groups)

D. Confirmation Dynamics

D.1. Approximations

First, some words on the behavior of $Z_{DOJ}(h, e)$. If 3.3, then the following approximation holds:

$$Z_{DOJ}(h, e) \approx DOJ(h|e) - DOJ(h) \tag{D.1}$$

If additionally 3.4 holds, then 3.2 follows.

Some words on the behavior of $F_{DOJ}(h, e)$ respectively $L_{DOJ}(h, e)$. Given 2.2 and 2.1, $L_{DOJ}(h, e)$ can be written as follows:

$$L_{DOJ}(h, e) = \frac{\frac{DOJ(h|e)}{DOJ(h)} - \frac{1-DOJ(h|e)}{1-DOJ(h)}}{\frac{DOJ(h|e)}{DOJ(h)} + \frac{1-DOJ(h|e)}{1-DOJ(h)}} \tag{D.2}$$

If 3.1 and 3.3, then 3.5 holds. It follows that $L_{DOJ}(h, e)$ increases with increasing $\frac{DOJ(h|e)}{DOJ(h)}$, compare Fig. D.1. Presuming further 3.4, $L_{DOJ}(h, e)$ can be fairly approximated as 1.

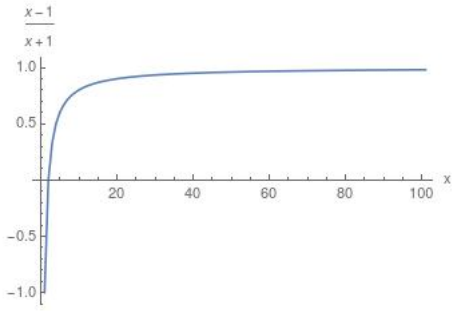

Figure D.1.: Illustration of 3.5, that is, some approximation of $L_{DOJ}(h, e)$ with $x = \frac{DOJ(h|e)}{DOJ(h)}$.

D.2. Additional Justification Results

The following three figures, that is Fig. D.2, Fig. D.3 and Fig. D.4, are sets of plots with every plot corresponding to a certain main participant, that is De la Beche (DLB), Murchison (MUR), Lyell (LYE), Phillips (PHI), Sedgwick (SED) or Austen (AUS). A single plot shows dynamics of the absolute degree of justification of a person's body of evidence ($DOJ(e)$), the absolute degree of justification of a person's dating hypothesis ($DOJ(h)$) and the ratio of the conditional degree of a person's dating hypothesis given her body of evidence and the absolute degree of justification of this very dating hypothesis ($\frac{DOJ(h|e)}{DOJ(h)}$), respectively.

From Fig. D.2, Fig. D.3 and Fig. D.4 follow 3.6, 3.3 and 3.4.

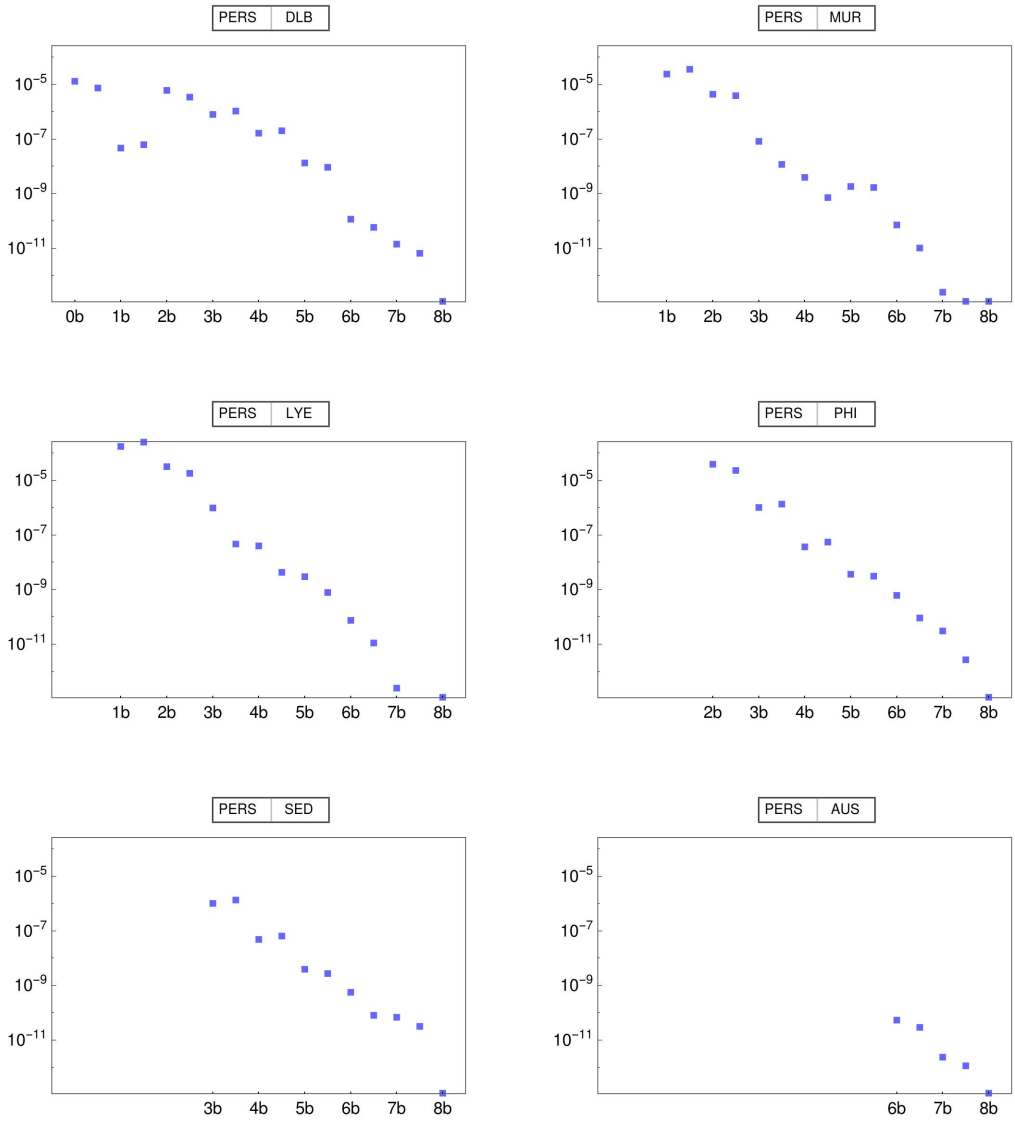

Figure D.2.: Dynamics of the absolute degree of justification of bodies of evidence. A set of plots is shown with every plot corresponding to a certain main participant, that is De la Beche (DLB), Murchison (MUR), Lyell (LYE), Phillips (PHI), Sedgwick (SED) or Austen (AUS). A single plot shows the absolute degree of a person's body of evidence ($DOJ(e)$) at different time steps.

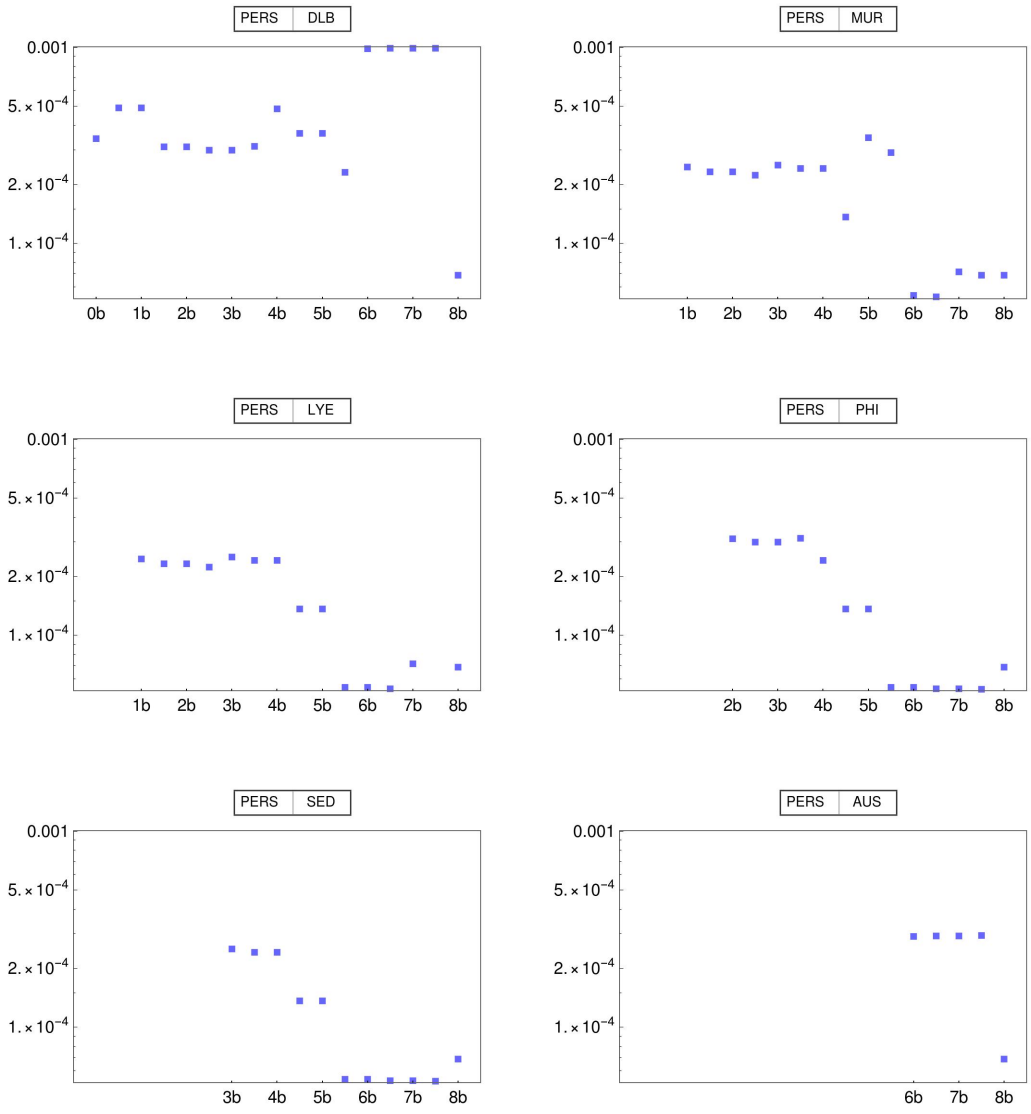

Figure D.3.: Dynamics of the absolute degree of justification of dating hypotheses. A set of plots is shown with every plot corresponding to a certain main participant, that is De la Beche (DLB), Murchison (MUR), Lyell (LYE), Phillips (PHI), Sedgwick (SED) or Austen (AUS). A single plot shows the absolute degree of a person's dating hypothesis ($DOJ(h)$) at different time steps.

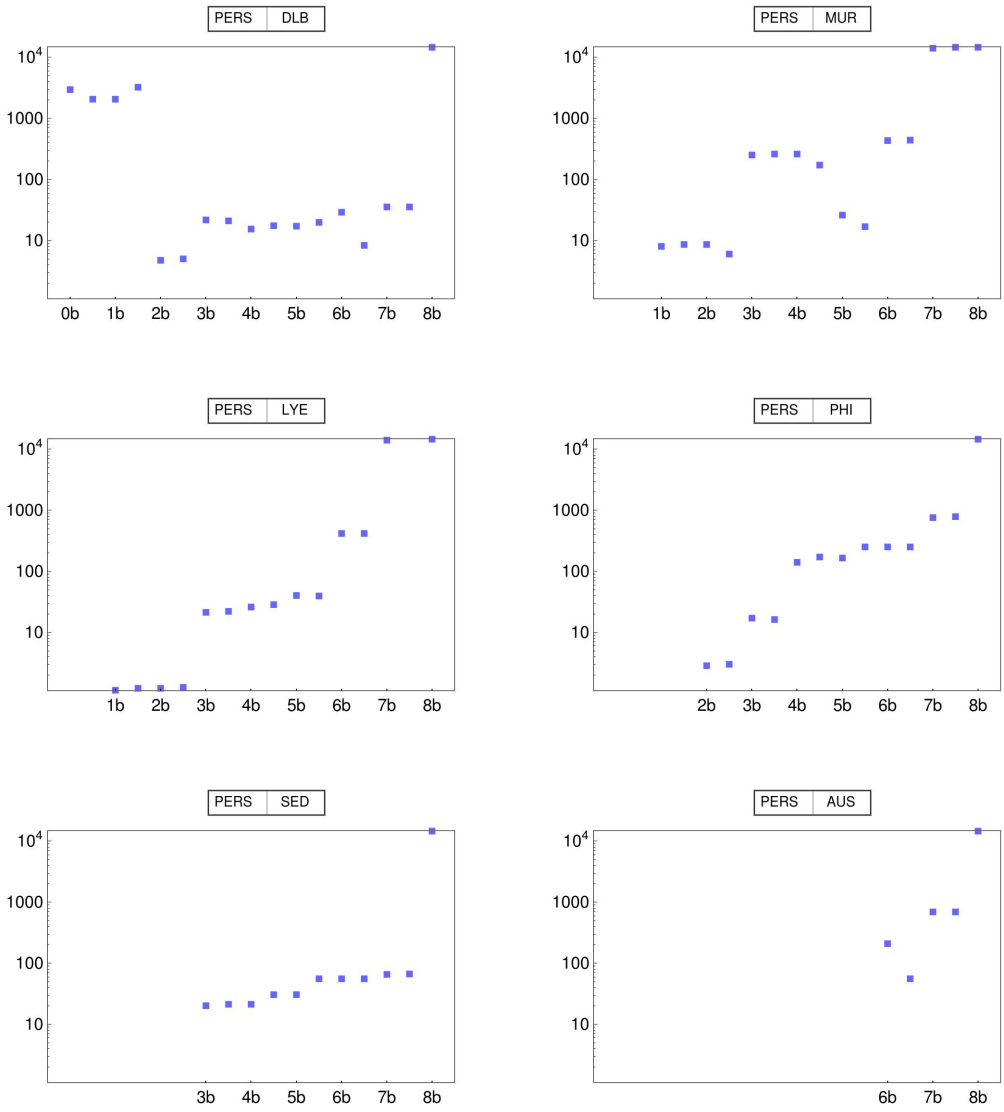

Figure D.4.: A set of plots is shown with every plot corresponding to a certain main participant, that is De la Beche (DLB), Murchison (MUR), Lyell (LYE), Phillips (PHI), Sedgwick (SED) or Austen (AUS). A single plot shows the ratio of the conditional degree of a person's dating hypothesis given her body of evidence and the absolute degree of justification of this very dating hypothesis ($\frac{DOJ(h|e)}{DOJ(h)}$) at different time steps

D.3. Possible Confirmation

For every time step, person and possible dating hypothesis, the degree to which the person's body of evidence confirms the dating hypothesis is calculated. Based on my reconstruction of the great Devonian controversy, there are 877 possible dating hypotheses.

The following plots show results for three different confirmation measures, namely $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$. Color and size of a pie sector correspond to a certain value of confirmation and its relative frequency, respectively. Turning from blue over yellow to red, confirmation increases.

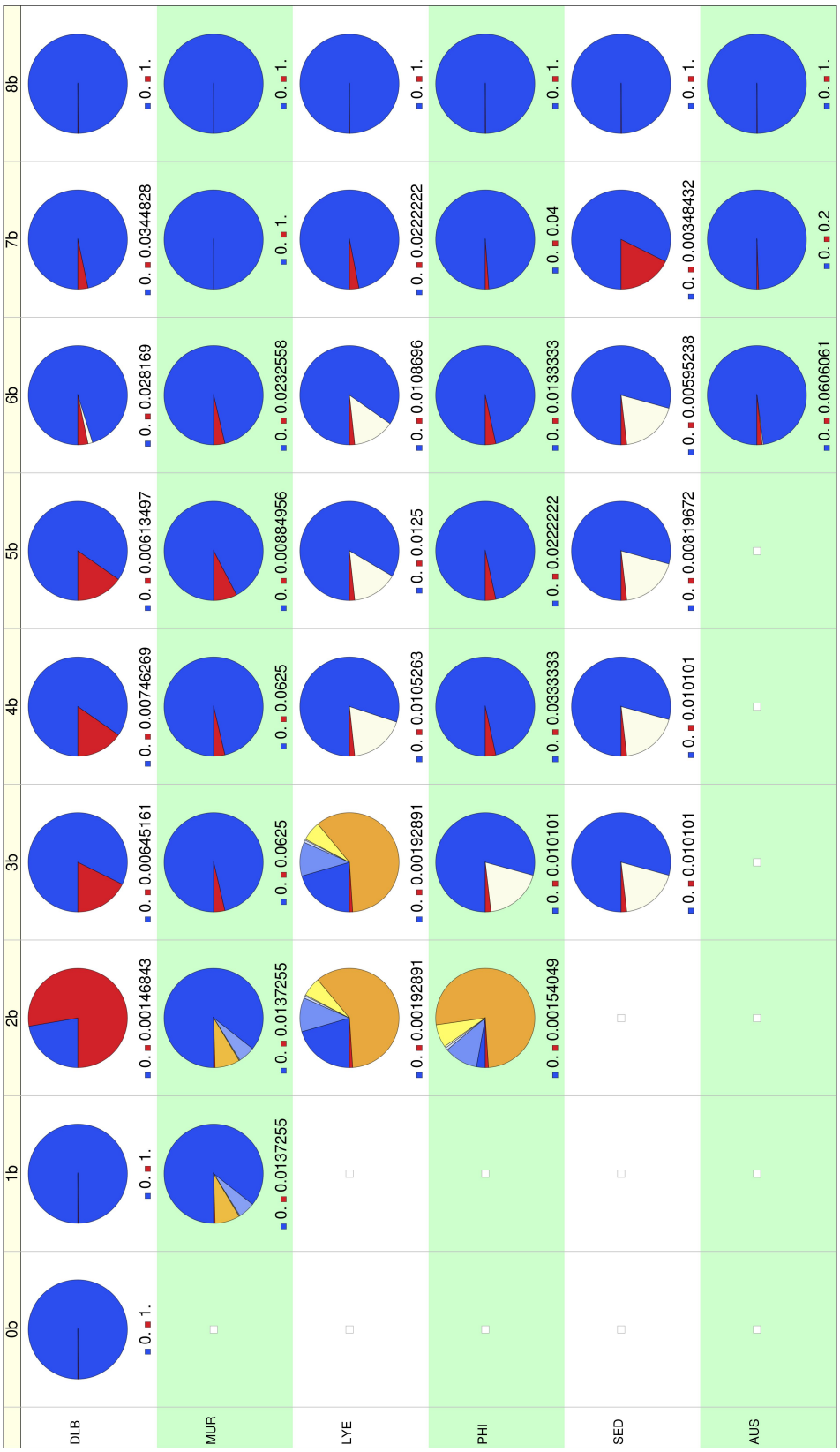

Figure D.5.: For every time step, t_b, person, j, and possible dating hypothesis, k, $DOJ(h_k|e_j^{t_b})$ is calculated. Based on my reconstruction of the debate, there are 877 possible dating hypotheses. Color and size of a pie sector correspond to the degree confirmation and its relative frequency, respectively. Turning from blue over yellow to red, confirmation increases.

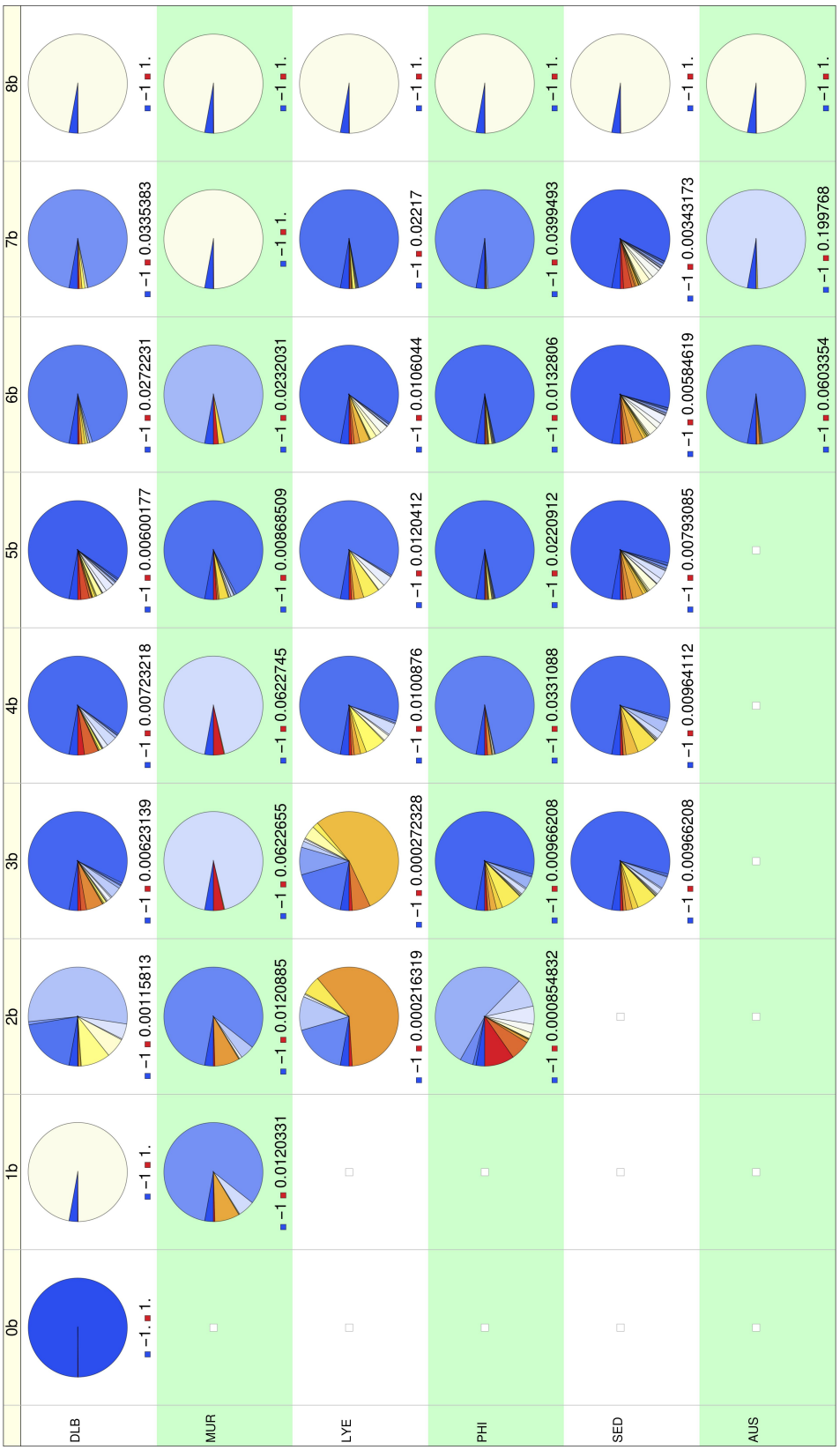

Figure D.6.: For every time step, t_b, person, j, and possible dating hypothesis, k, $Z_{DOJ}(h_k, e_j^{t_b})$ is calculated. Based on my reconstruction of the debate, there are 877 possible dating hypotheses. Color and size of a pie sector correspond to the degree confirmation and its relative frequency, respectively. Turning from blue over yellow to red, confirmation increases.

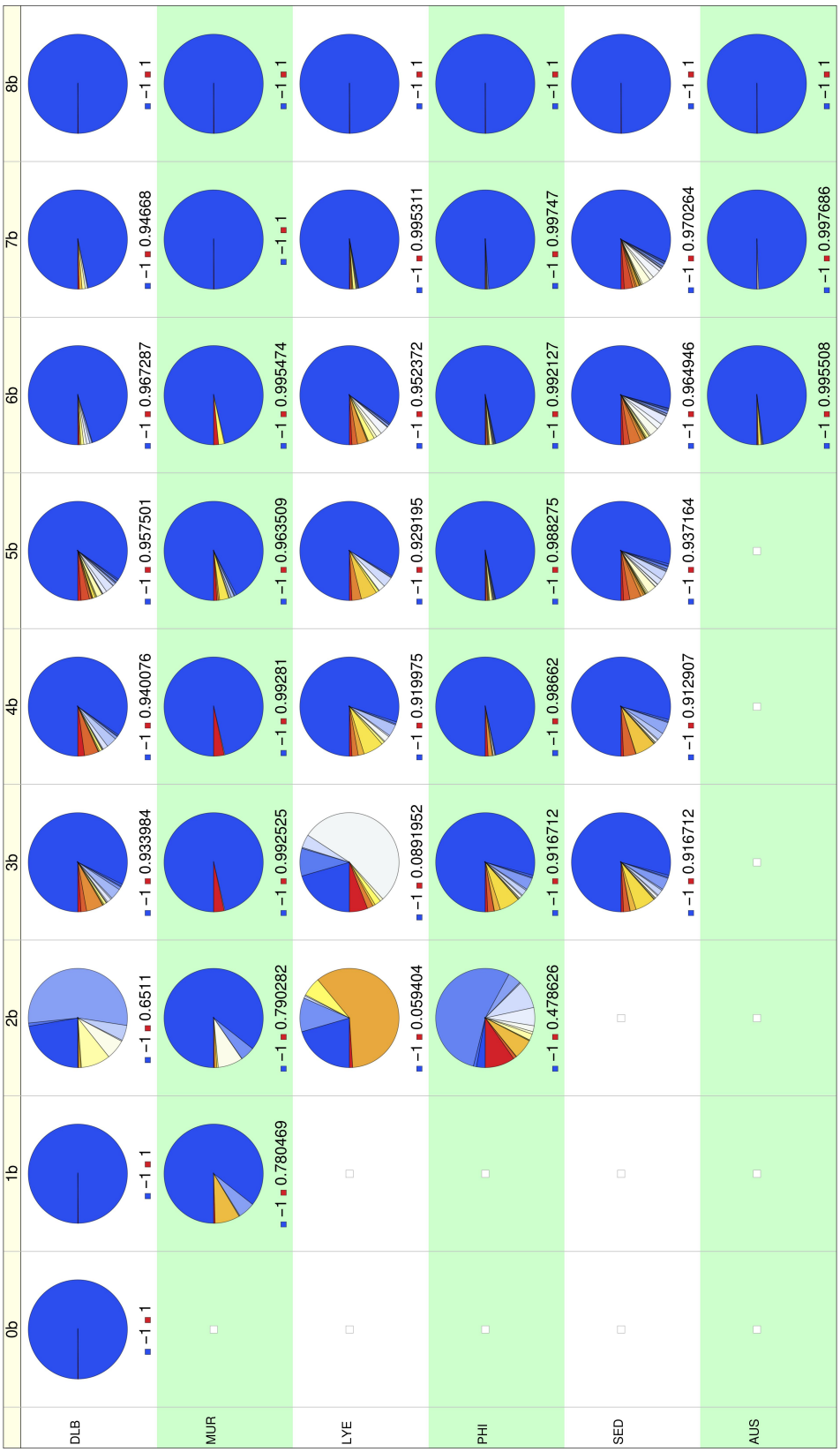

Figure D.7.: For every time step, t_b, person, j, and possible dating hypothesis, k, $F_{DOJ}(h_k, e_j^{t_b})$ is calculated. Based on my reconstruction of the debate, there are 877 possible dating hypotheses. Color and size of a pie sector correspond to the degree confirmation and its relative frequency, respectively. Turning from blue over yellow to red, confirmation increases.

D.4. Definitional Differences in Confirmation

Here, some differences between $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$ in maximizing confirmation are discussed. For simplicity, presume that $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$ are sufficiently well approximated by D.1 and 3.5, respectively. Further, presume that there is a dating hypothesis, h_i, maximizing confirmation:

$$\forall h_j \in H : DOJ(h_i|e) > DOJ(h_j|e) \text{ with } i \neq j \tag{D.3}$$

$$\forall h_j \in H : \frac{DOJ(h_i|e)}{DOJ(h_i)} > \frac{DOJ(h_j|e)}{DOJ(h_j)} \text{ with } i \neq j \tag{D.4}$$

$$\forall h_j \in H : DOJ(h_i|e) - DOJ(h_i) > DOJ(h_j|e) - DOJ(h_j) \text{ with } i \neq j \tag{D.5}$$

Here, H denotes the space of all possible dating hypotheses. Hence, there is a dating hypothesis, h_i, maximizing $Z_{DOJ}(h,e)$ respectively $F_{DOJ}(h,e)$, if there is no other dating hypothesis, h_j, with (i) a sufficiently large degree of conditional justification, $DOJ(h_j|e)$, and (ii) a sufficiently small degree of absolute justification, $DOJ(h_j)$.

Is there a dating hypothesis maximizing $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$ at the same time? What combinations are possible? Is there a dating hypothesis maximizing only one confirmation measure? Is there a dating hypothesis maximizing only two confirmation measures at the same time?

Let us presume that there is a dating hypothesis, h_i, maximizing respectively minimizing $DOJ(h_j)$:

$$\forall h_j \in H : DOJ(h_i) > DOJ(h_j) \text{ with } i \neq j \tag{D.6}$$

$$\forall h_j \in H : DOJ(h_i) < DOJ(h_j) \text{ with } i \neq j \tag{D.7}$$

D.6 and D.4 together imply D.3. D.7 and D.3 together imply D.4. Hence, for some dating hypothesis maximizing $DOJ(h_j)$, it holds: If it maximizes $F_{DOJ}(h_j, e)$, then it maximizes $DOJ(h_j|e)$, too. For some dating hypothesis minimizing $DOJ(h_j)$, it holds: If it maximizes $DOJ(h_j|e)$, then it maximizes $F_{DOJ}(h_j, e)$, too. Notice that, if some dating hypothesis does not maximize $F_{DOJ}(h_j, e)$ but $DOJ(h_j|e)$, then it does not minimize $DOJ(h_j)$. Notice also that, if some dating hypothesis does not maximize $DOJ(h_j|e)$ but $F_{DOJ}(h_j, e)$, then it does not maximize $DOJ(h_j)$.

D.6 and D.5 together imply D.3. D.7 and D.3 together imply D.5. Hence, for some dating hypothesis maximizing $DOJ(h_j)$, it holds: If it maximizes $Z_{DOJ}(h_j, e)$, then it maximizes $DOJ(h_j|e)$, too. For some dating hypothesis minimizing $DOJ(h_j)$, it holds: If it maximizes $DOJ(h_j|e)$, then it maximizes $Z_{DOJ}(h_j, e)$, too. Notice that, if some dating hypothesis does not maximize $Z_{DOJ}(h_j, e)$ but $DOJ(h_j|e)$, then it does not minimize $DOJ(h_j)$. Notice also that, if some dating hypothesis does not maximize $DOJ(h_j|e)$ but $Z_{DOJ}(h_j, e)$, then it does not maximize $DOJ(h_j)$.

However, dating hypotheses satisfying D.6 respectively D.7 need not be the only ones maximizing more than one confirmation measure.

D.5. Dating Hypotheses with a Maximal Degree of Confirmation

For all main participants, dating hypotheses with a maximal degree of confirmation are determined. However, results are shown only for those time steps at which the person accepts a dating hypothesis with a lower degree of confirmation. Additionally, every dating hypothesis is labeled with its similarity to the final consensus (CON_h).

It shows that sets of dating hypotheses *maximizing* confirmation are the same for relevance confirmation measures, the only exception being De la Beche at time step 6. Note that, at time step 3*b*, the set of dating hypotheses *maximizing* confirmation using a relevance confirmation measure is a subset of those for absolute confirmation, compare Fig. D.20 and Fig. D.19. It is generated by applying one additional condition, namely dating at least some part of the main part of the Culm as Coal Measures in age. Further, at the same time step, sets of dating hypotheses *maximizing* confirmation are the same for Lyell, Sedgwick and Phillips. The corresponding set for De la Beche differs in dating at least some part of the black Culm limestone

as old as the Non-Culm strata, compare once again Fig. D.20.

Figure D.8.: Dating hypotheses with a maximal degree of $F_{DOJ}(h, e)$ given De la Beche's body of evidence. However, only those time steps are shown at which De la Beche accepts some dating hypothesis with a lower degree of $F_{DOJ}(h, e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

Figure D.9.: Dating hypotheses with a maximal degree of $Z_{DOJ}(h, e)$ given De la Beche's body of evidence. However, only those time steps are shown at which De la Beche accepts some dating hypothesis with a lower degree of $Z_{DOJ}(h, e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

Figure D.10.: Dating hypotheses with a maximal degree of $DOJ(h|e)$ given Lyell's body of evidence. However, only those time steps are shown at which Lyell accepts some dating hypothesis with a lower degree of $DOJ(h|e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

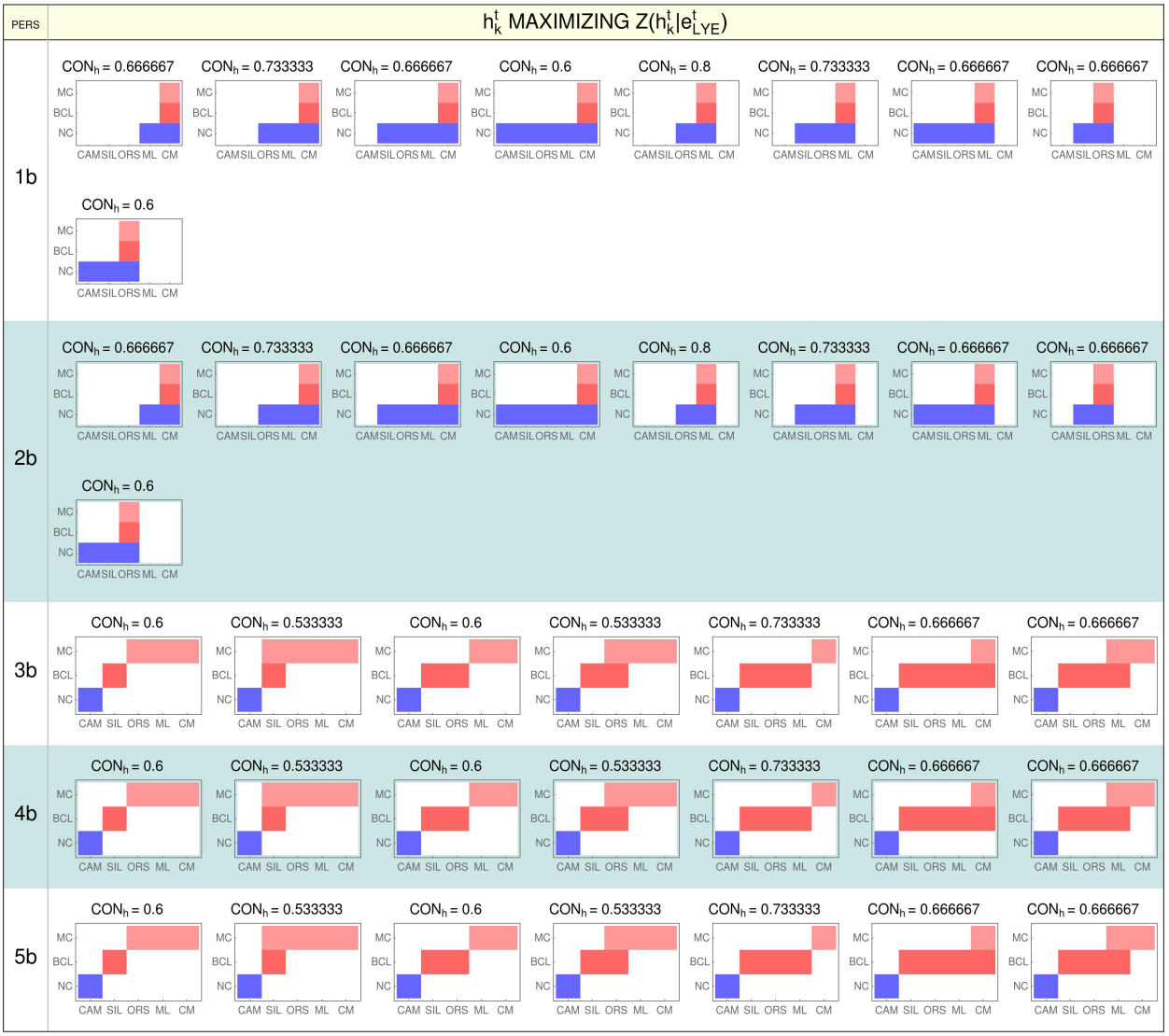

Figure D.11.: Dating hypotheses with a maximal degree of $Z_{DOJ}(h, e)$ given Lyell's body of evidence. However, only those time steps are shown at which Lyell accepts some dating hypothesis with a lower degree of $Z_{DOJ}(h, e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

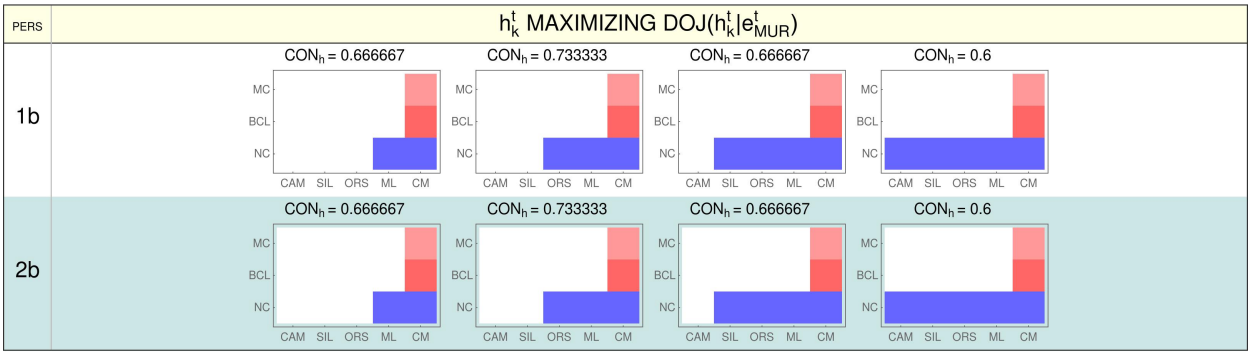

Figure D.12.: Dating hypotheses with a maximal degree of $DOJ(h|e)$ given Murchison's body of evidence. However, only those time steps are shown at which Murchison accepts some dating hypothesis with a lower degree of $DOJ(h|e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

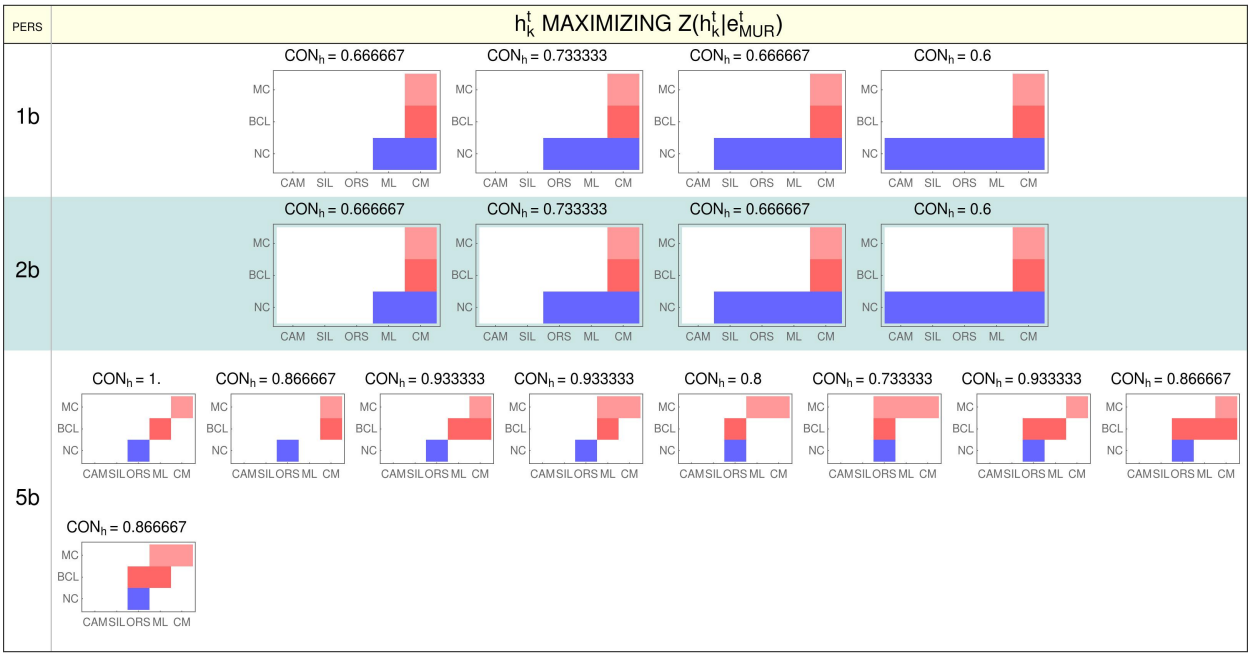

Figure D.13.: Dating hypotheses with a maximal degree of $Z_{DOJ}(h, e)$ given Murchison's body of evidence. However, only those time steps are shown at which Murchison accepts some dating hypothesis with a lower degree of $Z_{DOJ}(h, e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

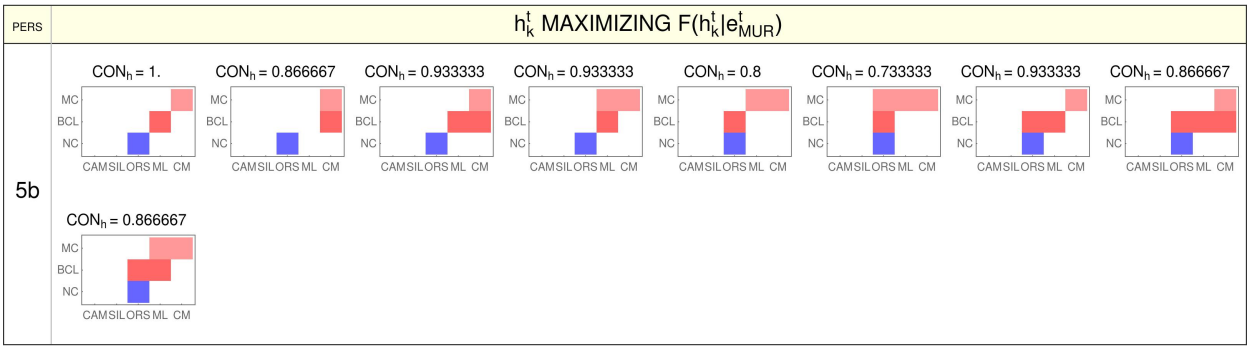

Figure D.14.: Dating hypotheses with a maximal degree of $F_{DOJ}(h, e)$ given Murchison's body of evidence. However, only those time steps are shown at which Murchison accepts some dating hypothesis with a lower degree of $F_{DOJ}(h, e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

Figure D.15.: Dating hypotheses with a maximal degree of $DOJ(h|e)$ given Sedgwick's body of evidence. However, only those time steps are shown at which Sedgwick accepts some dating hypothesis with a lower degree of $DOJ(h|e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

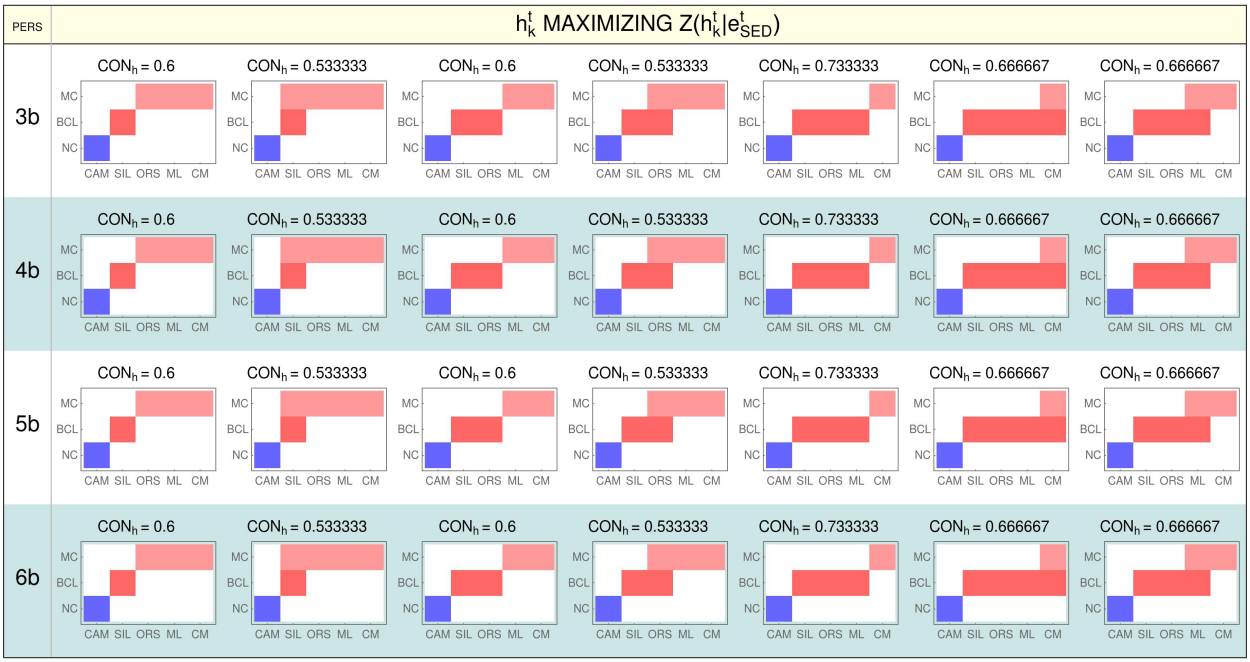

Figure D.16.: Dating hypotheses with a maximal degree of $Z_{DOJ}(h, e)$ given Sedgwick's body of evidence. However, only those time steps are shown at which Sedgwick accepts some dating hypothesis with a lower degree of $Z_{DOJ}(h, e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

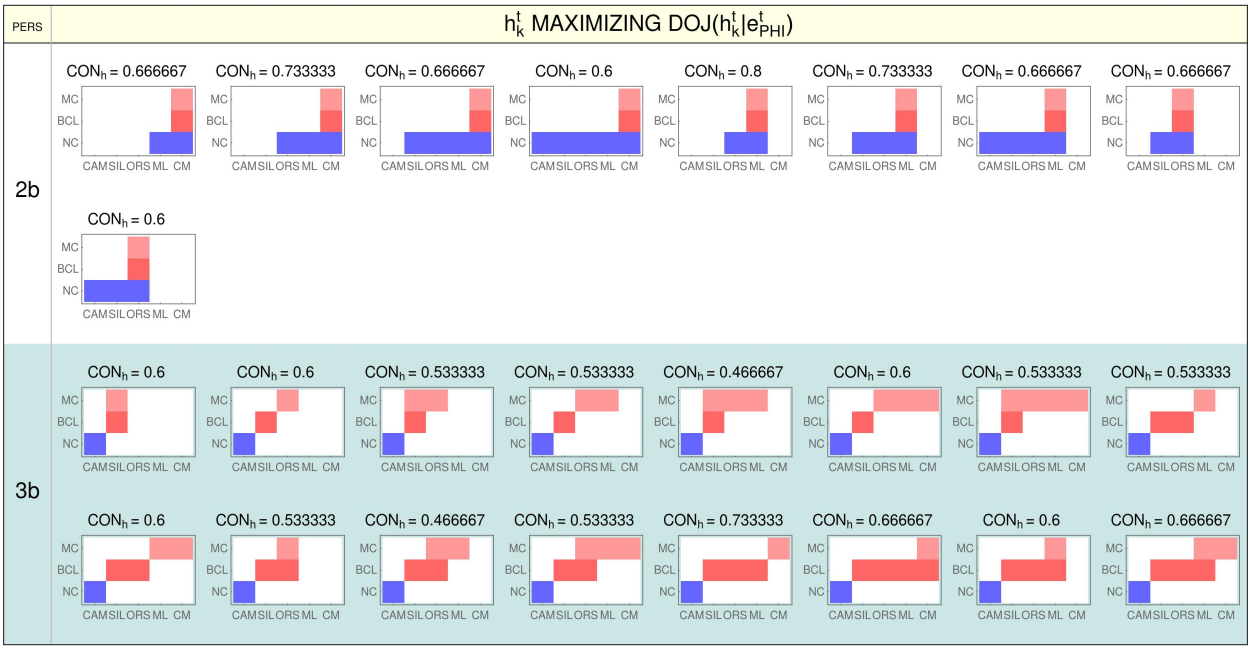

Figure D.17.: Dating hypotheses with a maximal degree of $DOJ(h|e)$ given Phillips's body of evidence. However, only those time steps are shown at which Phillips accepts some dating hypothesis with a lower degree of $DOJ(h|e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

Figure D.18.: Dating hypotheses with a maximal degree of $Z_{DOJ}(h, e)$ given Phillips's body of evidence. However, only those time steps are shown at which Phillips accepts some dating hypothesis with a lower degree of $Z_{DOJ}(h, e)$. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

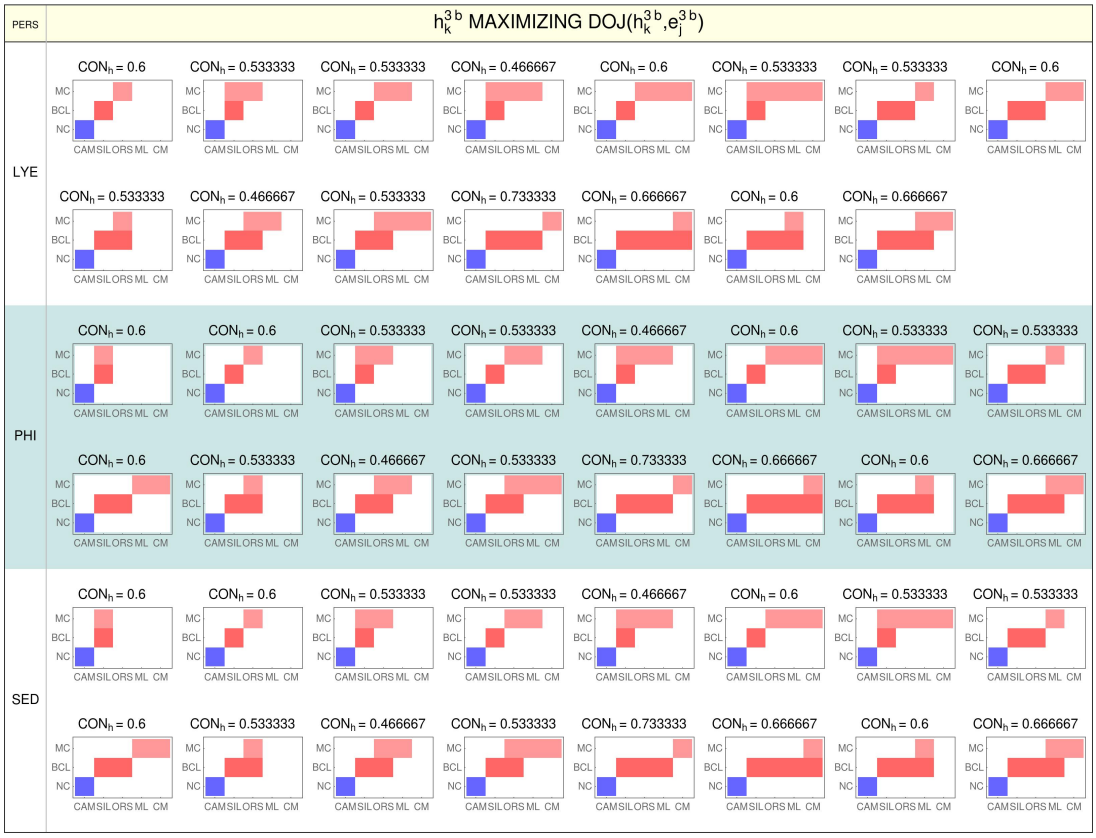

Figure D.19.: At time step $3b$, dating hypotheses with a maximal degree of $DOJ(h|e)$ given a person's body of evidence. However, only those persons are shown which accept some dating hypothesis with a lower degree of $DOJ(h|e)$ at that time. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

Figure D.20.: At time step $3b$, dating hypotheses with a maximal degree of $Z_{DOJ}(h, e)$ given a person's body of evidence. However, only those persons are shown which accept some dating hypothesis with a lower degree of $Z_{DOJ}(h, e)$ at that time. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

Figure D.21.: At time step $3b$, dating hypotheses with a maximal degree of $F_{DOJ}(h, e)$ given a person's body of evidence. However, only those persons are shown which accept some dating hypothesis with a lower degree of $F_{DOJ}(h, e)$ at that time. Additionally, every dating hypothesis is labeled with its similarity to the final consensus. Additionally, every dating hypothesis is labeled with its similarity to the final consensus.

E. Roads to the Final Consensus

E.1. Similarity with the Final Consensus

Time	Non–Empirical Statements	Empirical Statements
S8	[LV in Sedimentation] [Main Culm Youngest Devonian Strata] [Characteristic Fossil Assemblage Principle – V2] [Non–Culm – Body of Evidence – Fossils] ! [Non–Culm – Body of Evidence – Region] [CFA – ORS – III – V2]	[Philipps' Collection] ! [Sequence of Strata – Tor Bay and Newton Abott] [Carboniferous Plants in North Devon] [Scottish ORS – Fossils] ! [Passing Devon Northwards]

Figure E.1.: Shared beliefs at the final time step. Here, only sentence titles are shown. Corresponding sentences are listed in Fig. B.12 and Fig. B.13.

Time	DLB	MUR	LYE	PHI	SED	AUS
S0b	0.6					
S1a	0.6					
S1b	0.6	0.73333	0.73333			
S2a	0.6	0.73333	0.73333			
S2b	0.6	0.73333	0.73333	0.6		
S3a	0.6	0.73333	0.73333	0.6		
S3b	0.6	0.66667	0.66667	0.6	0.66667	
S4a	0.6	0.66667	0.66667	0.6	0.66667	
S4b	0.53333	0.66667	0.66667	0.8	0.8	
S5a	0.53333	0.66667	0.66667	0.8	0.8	
S5b	0.53333	0.73333	0.66667	0.8	0.8	
S6a	0.53333	0.73333	0.66667	0.8	0.8	
S6b	0.6	0.8	0.8	0.8	0.66667	0.86667
S7a	0.6	0.8	0.8	0.8	0.66667	0.86667
S7b	0.6	1.	1.	0.8	0.66667	0.86667
S8a	0.6	1.	1.	0.8	0.66667	0.86667
S8b	1.	1.	1.	1.	1.	1.

Figure E.2.: Similarity with the final dating hypothesis. For every person and time step, the similarity between the person's dating hypothesis and the final consensus dating hypothesis is determined.

Time	DLB	MUR	LYE	PHI	SED	AUS
S0b	0.51923					
S1a	0.42308					
S1b	0.53333	0.70833	0.66667			
S2a	0.6	0.71681	0.69643			
S2b	0.77143	0.78049	0.7561	0.86066		
S3a	0.77551	0.7907	0.76154	0.86047		
S3b	0.93878	0.84507	0.82394	0.91489	0.91489	
S4a	0.90698	0.77528	0.79213	0.88554	0.88554	
S4b	0.9116	0.76667	0.78333	0.90857	0.8453	
S5a	0.87958	0.72727	0.75	0.87568	0.81675	
S5b	0.88945	0.79188	0.72772	0.88601	0.8209	
S6a	0.88152	0.78947	0.72897	0.87805	0.8169	
S6b	0.80717	0.70909	0.72936	0.86124	0.80184	0.9148
S7a	0.77872	0.69828	0.71739	0.84163	0.78603	0.90129
S7b	0.76569	0.93443	0.84146	0.84	0.80335	0.89121
S8a	0.76426	0.91111		0.8419	0.78161	0.87833
S8b	0.9631	0.97026	0.96667	0.9631	0.9631	0.9631

Figure E.3.: Similarity with the final consensus body of evidence. For every person and time step, the similarity between the person's body of evidence and the final consensus body of evidence is determined.

S,A,C						
Time	DLB	MUR	LYE	PHI	SED	AUS
S0b	{1, 9, 8}					
S1a	{6, 11, 13}					
S1b	{3, 16, 13}	{15, 17, 2}	{15, 16, 3}			
S2a	{3, 27, 17}	{23, 27, 3}	{22, 26, 4}			
S2b	{8, 36, 8}	{18, 32, 3}	{18, 31, 4}	{17, 35, 0}		
S3a	{9, 38, 8}	{18, 34, 3}	{19, 33, 4}	{18, 37, 0}		
S3b	{9, 46, 0}	{13, 40, 3}	{13, 39, 4}	{12, 43, 0}	{12, 43, 0}	
S4a	{16, 52, 0}	{16, 46, 8}	{16, 47, 7}	{19, 49, 0}	{19, 49, 0}	
S4b	{13, 55, 1}	{15, 46, 9}	{15, 47, 8}	{16, 53, 0}	{13, 51, 5}	
S5a	{20, 56, 1}	{18, 48, 12}	{19, 49, 10}	{23, 54, 0}	{20, 52, 5}	
S5b	{16, 59, 2}	{17, 52, 8}	{16, 49, 13}	{19, 57, 1}	{15, 55, 7}	
S6a	{19, 62, 2}	{20, 55, 8}	{19, 52, 13}	{22, 60, 1}	{18, 58, 7}	
S6b	{13, 60, 10}	{16, 52, 16}	{17, 53, 14}	{20, 60, 3}	{16, 58, 9}	{13, 68, 2}
S7a	{16, 61, 12}	{19, 54, 17}	{20, 55, 15}	{23, 62, 4}	{19, 60, 10}	{17, 70, 2}
S7b	{14, 61, 14}	{10, 76, 2}	{12, 69, 9}	{21, 63, 5}	{14, 64, 11}	{14, 71, 4}
S8a	{14, 67, 16}	{9, 82, 5}	{87, 0, 0}	{19, 71, 7}	{15, 68, 14}	{14, 77, 6}
S8b	{10, 87, 0}	{8, 87, 0}	{9, 87, 0}	{10, 87, 0}	{10, 87, 0}	{10, 87, 0}

Figure E.4.: Similarity with the final consensus body of evidence depends on the number of judgment suspensions, agreements, and contradictions with the final consensus referred to as S, A and C, respectively. For every person and time step, these are listed.

E.2. Maximal Confirmation and Similarity with the Final Consensus

E.2.1. Similarity of Maximally Confirmed Hypotheses

For every time step, every person and every dating hypothesis which is maximally confirmed by the person's body of evidence at that time, the similarity between the dating hypothesis and the final dating hypothesis is calculated. Color and size of a pie sector correspond to the similarity value and its relative frequency, respectively. Turning from blue over yellow to red, similarity with the final dating hypothesis increases. On top of each pie, there is the size of the set of dating hypotheses which are maximally confirmed by the person's body of evidence at that time.

The following three plots correspond to results using $DOJ(h|e)$, $Z_{DOJ}(h, e)$ and $F_{DOJ}(h, e)$.

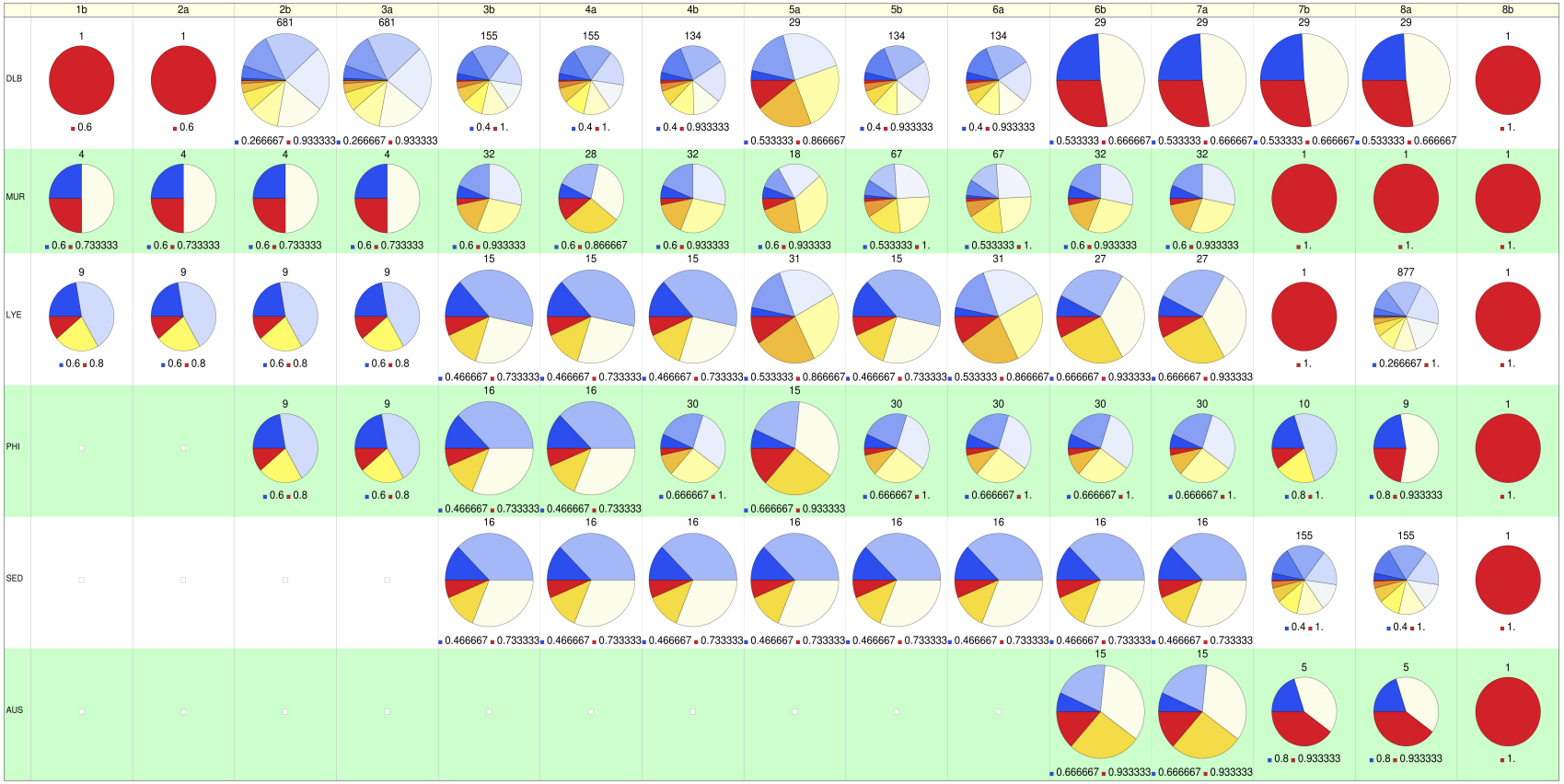

Figure E.5.: For every time step, every person and every dating hypothesis whith a maximal value of $DOJ(h|e)$ given the person's body of evidence, similarity with the final consensus is shown. Color and size of a pie sector correspond to the similarity value and its relative frequency, respectively. Turning from blue over yellow to red, similarity increases. On top of each pie, there is the size of the set of dating hypotheses whith a maximal value of $DOJ(h|e)$ given the person's body of evidence.

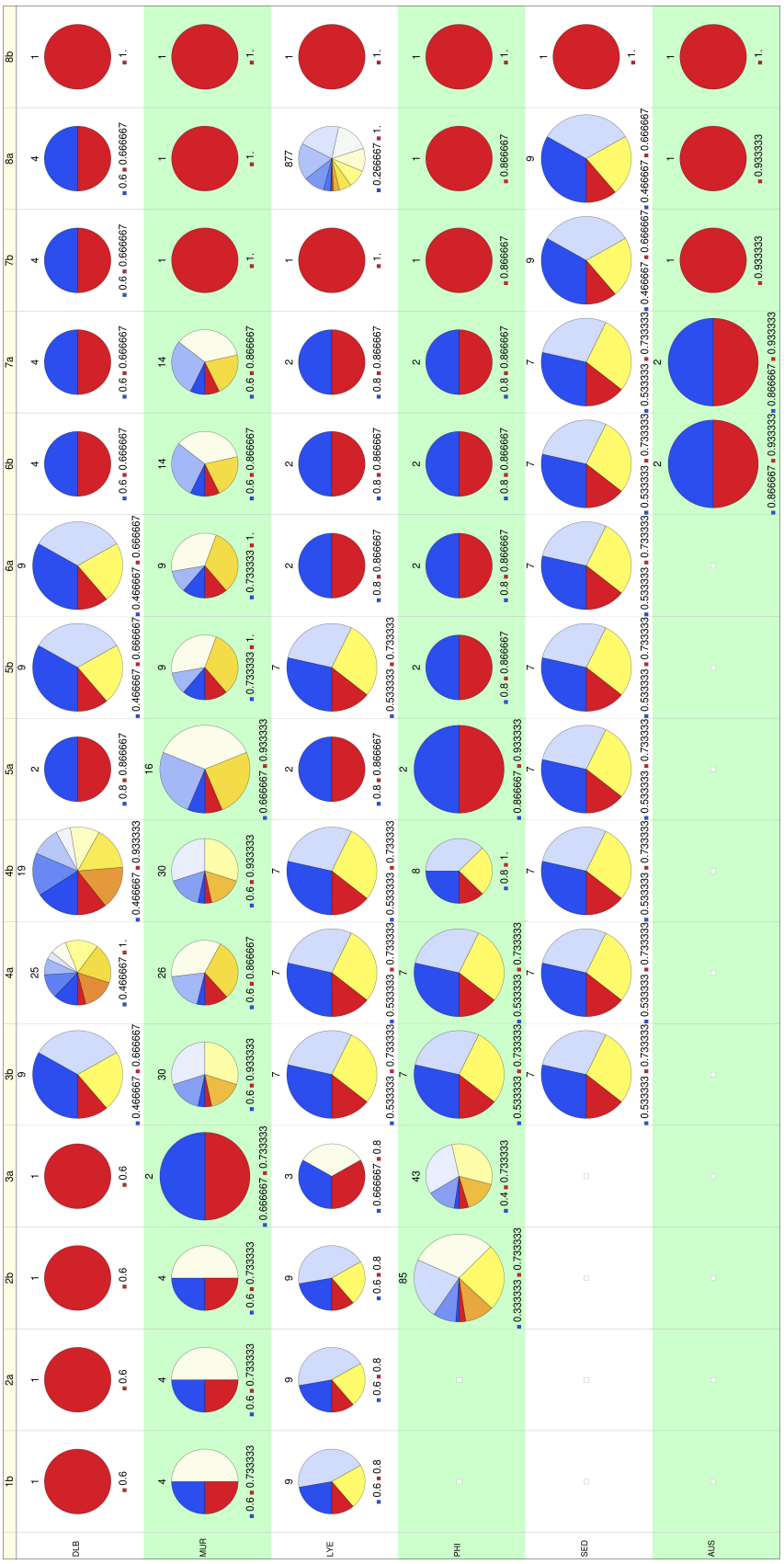

Figure E.6.: For every time step, every person and every dating hypothesis whith a maximal value of $Z_{DOJ}(h, e)$ given the person's body of evidence, similarity with the final consensus is shown. Color and size of a pie sector correspond to the similarity value and its relative frequency, respectively. Turning from blue over yellow to red, similarity increases. On top of each pie, there is the size of the set of dating hypotheses whith a maximal value of $Z_{DOJ}(h, e)$ given the person's body of evidence.

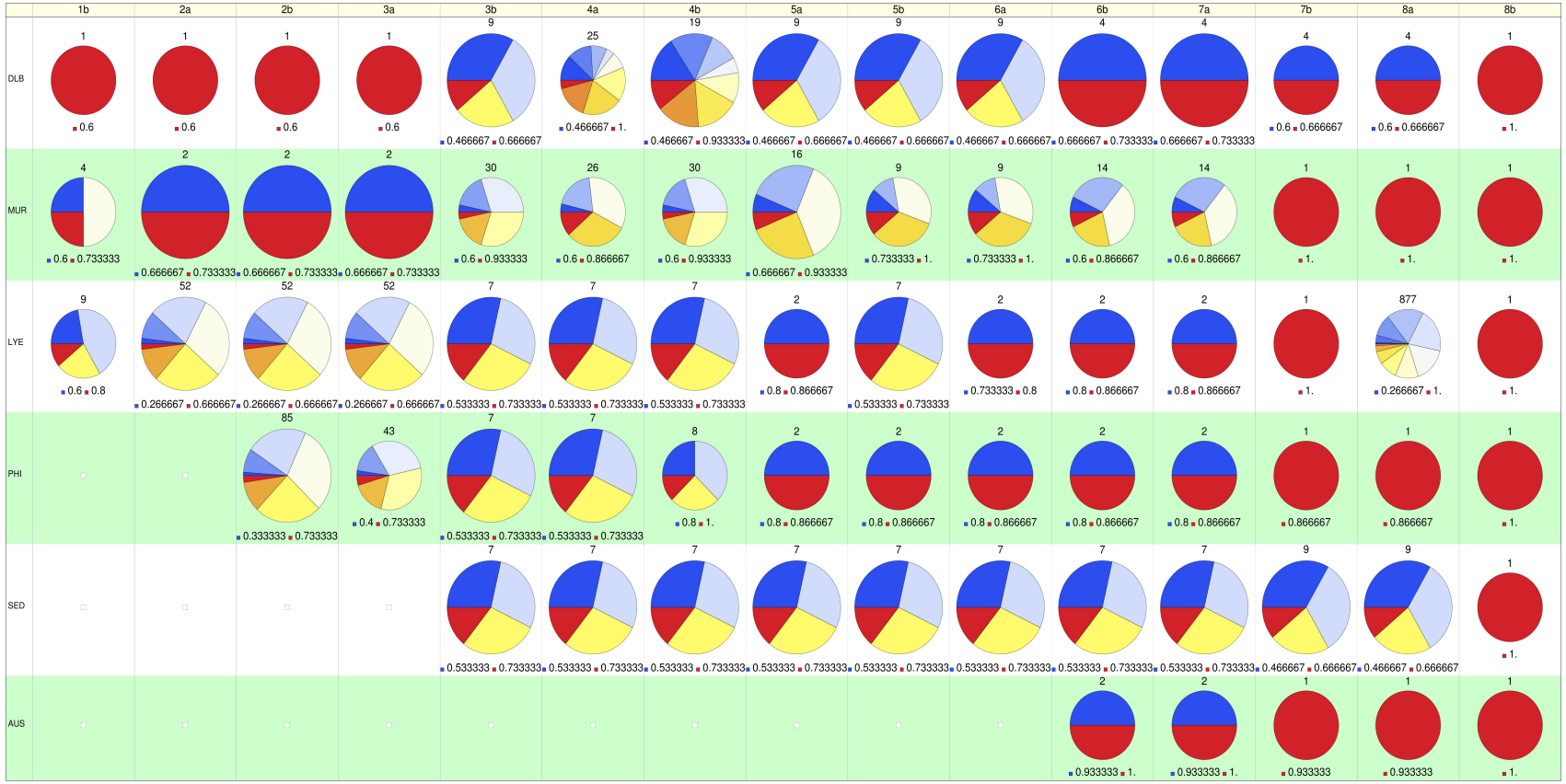

Figure E.7.: For every time step, every person and every dating hypothesis whith a maximal value of $F_{DOJ}(h, e)$ given the person's body of evidence, similarity with the final consensus is shown. Color and size of a pie sector correspond to the similarity value and its relative frequency, respectively. Turning from blue over yellow to red, similarity increases. On top of each pie, there is the size of the set of dating hypotheses whith a maximal value of $F_{DOJ}(h, e)$ given the person's body of evidence.

E.2.2. Sufficiently Similar and Maximally Confirmed Hypotheses

As in the previous subsection, for every time step, every person and every dating hypothesis which is maximally confirmed by the person's body of evidence at that time, the similarity between the dating hypothesis and the final dating hypothesis is shown. In contrast to the previous plots, only those dating hypotheses are considered, which are sufficiently similar to the final consensus, that is results are based on a subset of the previous set of data.

Color and size of a pie sector correspond to the similarity value and its relative frequency, respectively. Turning from blue over yellow to red, similarity with the final dating hypothesis increases. On top of each pie, there is the size of the set of dating hypotheses which are maximally confirmed by the person's body of evidence at that time and which are additionally sufficiently similar to the final consensus.

The following three plots correspond to results using $DOJ(h|e)$, $Z_{DOJ}(h, e)$ and $F_{DOJ}(h, e)$.

Figure E.8.: For every time step, every person and every dating hypothesis with a maximal value of $DOJ(h|e)$, given the person's body of evidence, and a sufficiently high similartiy with the final consensus, similarity with the final consensus is shown. Turning from blue over yellow to red, similarity increases. On top of each pie, there is the size of the set of dating hypotheses whith a maximal value of $DOJ(h|e)$ given the person's body of evidence and a sufficiently high similarity with the final consensus.

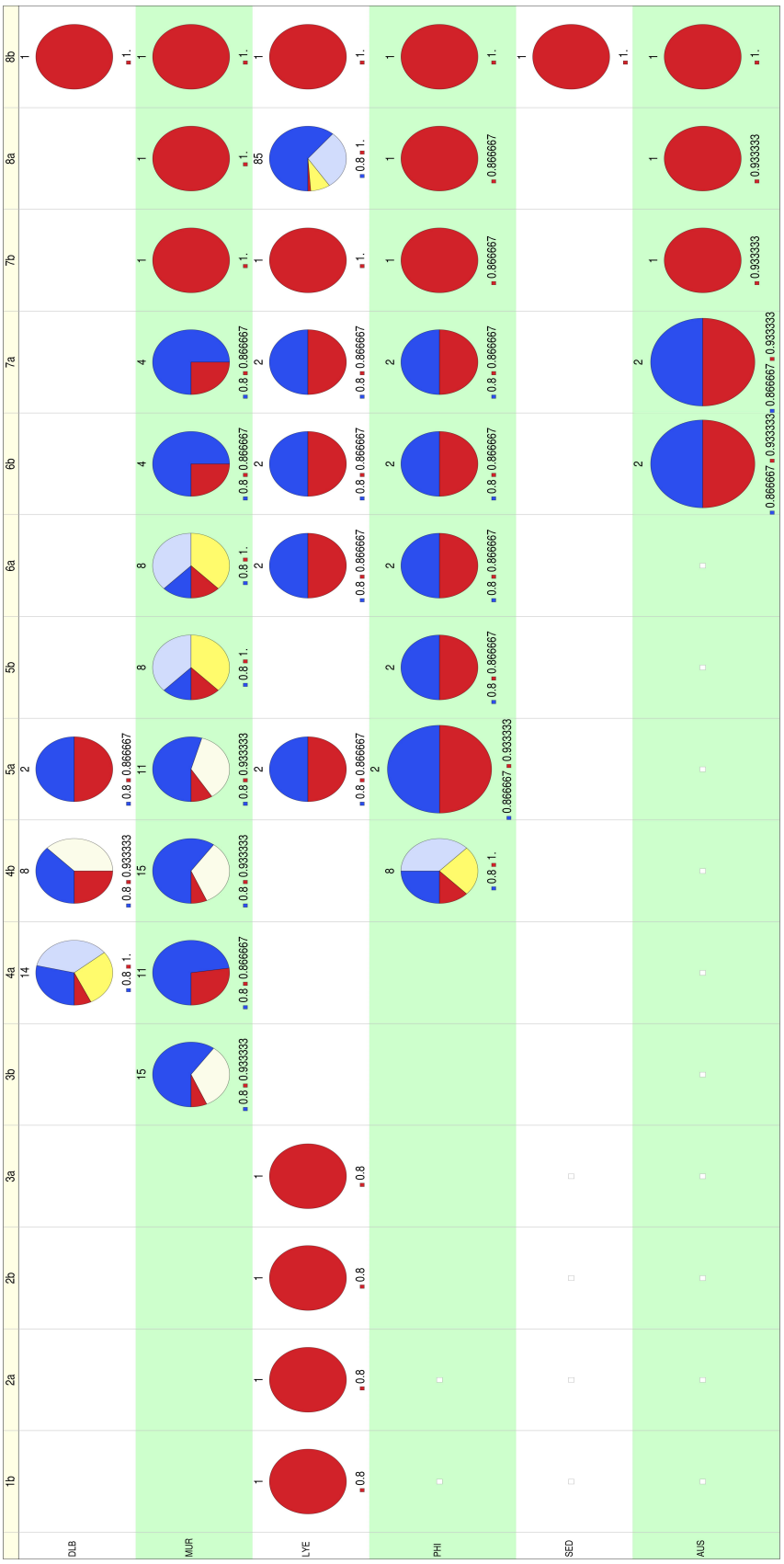

Figure E.9.: For every time step, every person and every dating hypothesis with a maximal value of $Z_{DOJ}(h, e)$, given the person's body of evidence, and a sufficiently high similartiy with the final consensus, similarity with the final consensus is shown. Turning from blue over yellow to red, similarity increases. On top of each pie, there is the size of the set of dating hypotheses whith a maximal value of $Z_{DOJ}(h, e)$ given the person's body of evidence and a sufficiently high similarity with the final consensus.

Figure E.10.: For every time step, every person and every dating hypothesis with a maximal value of $F_{DOJ}(h, e)$, given the person's body of evidence, and a sufficiently high similartiy with the final consensus, similarity with the final consensus is shown. Turning from blue over yellow to red, similarity increases. On top of each pie, there is the size of the set of dating hypotheses whith a maximal value of $F_{DOJ}(h, e)$ given the person's body of evidence and a sufficiently high similarity with the final consensus.

E.2.3. Success Ratio, Agreements and Contradictions with the Final Consensus

Is there a dependence between the success ratio and the number of agreements and contradictions of a body of evidence with the final consensus? The success ratio is introduced in sec. 3.4.3 as the ratio of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize confirmation, given a person's body of evidence at a certain time step.

For every person and time step, agreements are plotted over contradictions between the person's body of evidence and the deductive closure of the final consensus body of evidence. Additionally, those time steps are marked with a red circle where the ratio of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize confirmation is greater than 0.5. The size of the red circle relates to the actual value of this ratio. The following three plots, namely Fig. E.11, Fig. E.12 and Fig. E.13, correspond to results using $DOJ(h|e)$, $Z_{DOJ}(h,e)$ and $F_{DOJ}(h,e)$ as confirmation measures.

It shows that, for all three confirmation measures, it does not hold that the success ratio is greater 0.5 iff the number of agreements and contradictions are sufficiently large and small, respectively. Take as an example, Lyell at time step $6a$ using $Z_{DOJ}(h,e)$. Here, the number of contradictions is comparatively large, but the success ratio is greater 0.5. Take as another example De la Beche at time step $5b$ and $6a$ using $Z_{DOJ}(h,e)$. Here, the number of agreements and contradictions are of the same order as those of Phillips at time step $6b$. However, only for the latter, it holds that the success ratio is greater 0.5.

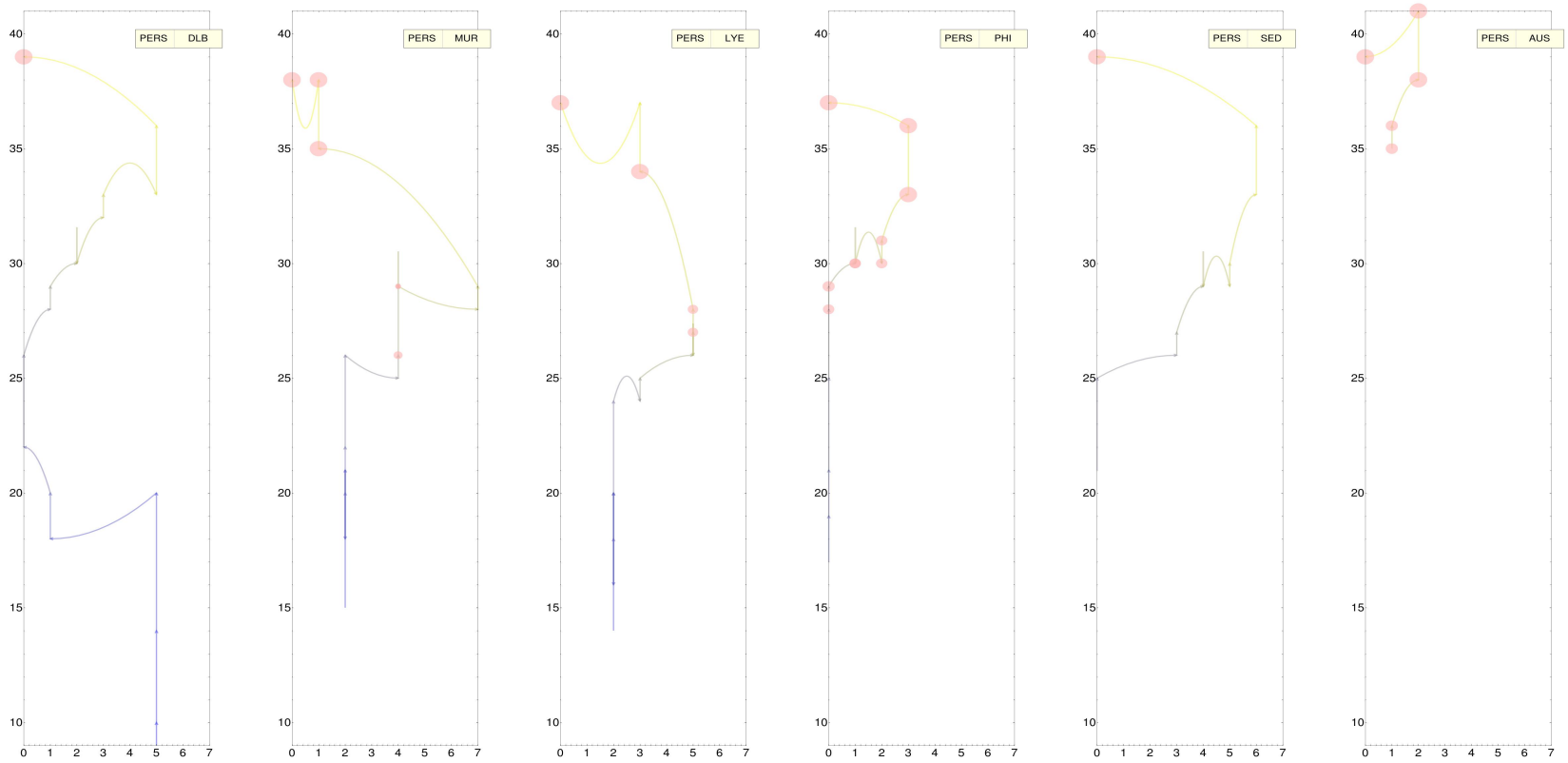

Figure E.11.: For every person and time step, agreements are plotted over contradictions between the person's body of evidence and the deductive closure of the final consensus body of evidence. Additionally, those time steps are marked with a red circle where the ratio of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize $DOJ(h|e)$ is greater than 0.5. The size of the red circle relates to the actual value of this ratio.

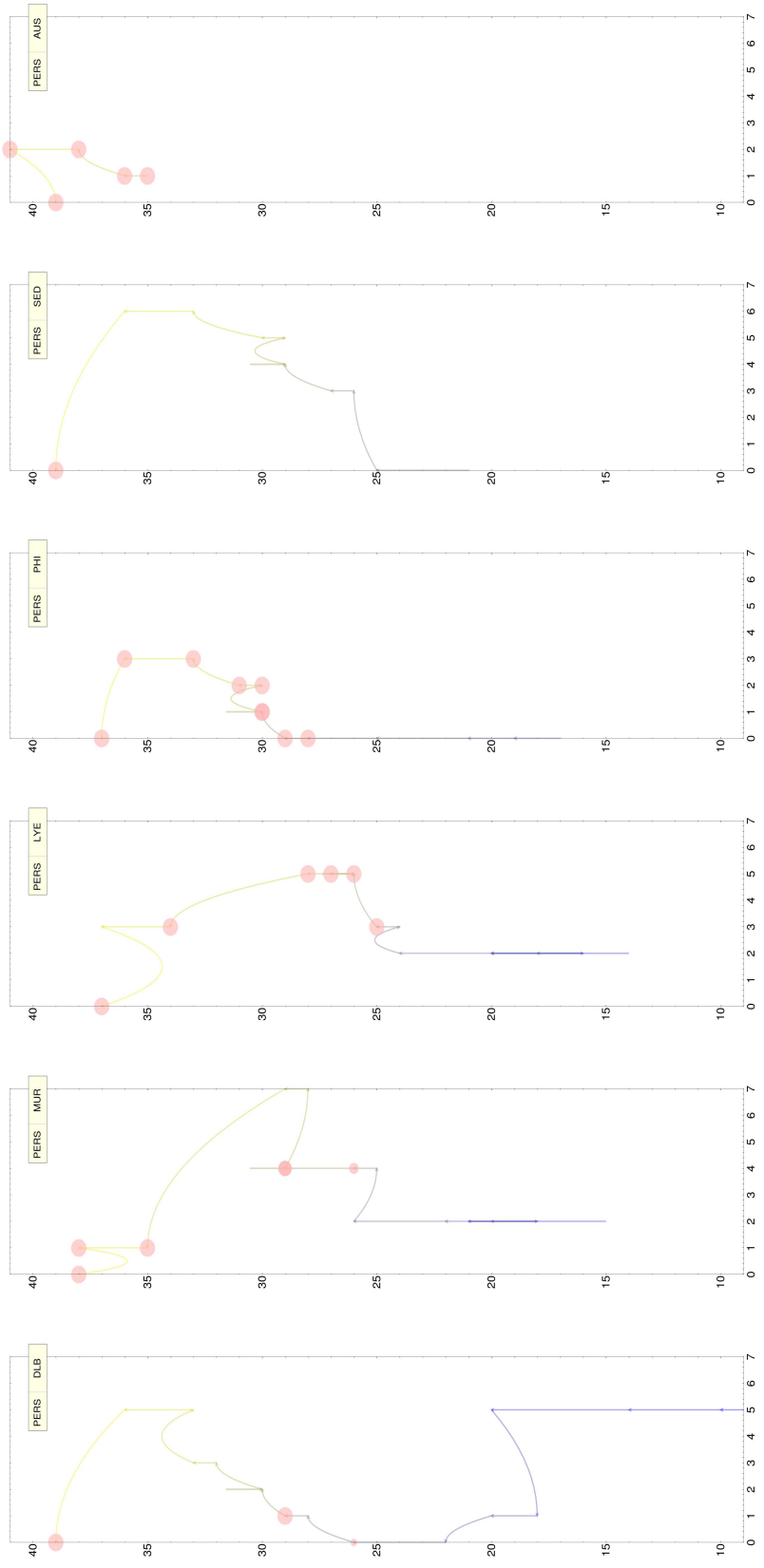

Figure E.12.: For every person and time step, agreements are plotted over contradictions between the person's body of evidence and the deductive closure of the final consensus body of evidence. Additionally, those time steps are marked with a red circle where the ratio of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize $Z_{DOJ}(h, e)$ is greater than 0.5. The size of the red circle relates to the actual value of this ratio.

Figure E.13.: For every person and time step, agreements are plotted over contradictions between the person's body of evidence and the deductive closure of the final consensus body of evidence. Additionally, those time steps are marked with a red circle where the ratio of dating hypotheses which are sufficiently similar to the final dating hypothesis among all those which maximize $F_{DOJ}(h, e)$ is greater than 0.5. The size of the red circle relates to the actual value of this ratio.

Bibliography

Claus Beisbart and John D. Norton. Why monte carlo simulations are inferences and not experiments. *International Studies in the Philosophy of Science*, 26(4): 403–422, 2012. doi: 10.1080/02698595.2012.748497.

Gregor Betz. *Theorie dialektischer Strukturen*. Klostermann, Frankfurt am Main, 2010. ISBN 978-3-465-03629-6.

Gregor Betz. On Degrees of Justification. *Erkenntnis*, 77(2):237–272, September 2012. ISSN 0165-0106, 1572-8420. doi: 10.1007/s10670-011-9314-y. URL `http://link.springer.com/10.1007/s10670-011-9314-y`.

Gregor Betz. *Debate dynamics how controversy improves our beliefs*. Springer, Dordrecht; New York, 2013. ISBN 978-94-007-4599-5 978-94-007-4598-8.

Gregor Betz. Truth in Evidence and Truth in Arguments without Logical Omniscience. *The British Journal for the Philosophy of Science*, page axv015, 2015. ISSN 0007-0882, 1464-3537. doi: 10.1093/bjps/axv015. URL `http://bjps.oxfordjournals.org/content/early/2015/03/31/bjps.axv015`.

Nora Mills Boyd and James Bogen. Theory and Observation in Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

Aaron Bramson, Patrick Grim, Daniel J. Singer, William J. Berger, Graham Sack, Steven Fisher, Carissa Flocken, and Bennett Holman. Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science*, 84(1):115–159, 1 2017. ISSN 0031-8248. doi: 10.1086/688938.

Branden Fitelson. Two Technical Corrections to My Coherence Measure. http://fitelson.org/coherence2.pdf, 2004. Online; accessed 23 June 2021.

Rudolf Carnap. Logical foundations of probability. *Mind*, 62(245):86–99, 1950.

Alan F. Chalmers. *Wege der Wissenschaft: EinfÃŒhrung in die Wissenschaftstheorie*. Springer, Berlin Heidelberg New York, 6., verb. aufl edition, 2007. ISBN 978-3-540-49490-4.

David Christensen. Does murphy's law apply in epistemology? *Oxford Studies in Epistemology*, 2:3–31, 2008.

David Christensen. Higher-order evidence. *Philosophy and Phenomenological Research*, 81(1):185–215, 2010.

Christian Voigt. Argdown - A simple syntax for complex argumentation. `https://argdown.org/`, 2018. Online; accessed 23 June 2021.

Vincenzo Crupi. Confirmation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

Vincenzo Crupi and Katya Tentori. Irrelevant Conjunction: Statement and Solution of a New Paradox. *Philosophy of Science*, 77(1):1–13, January 2010. ISSN 0031-8248. doi: 10.1086/650205. URL `http://www.journals.uchicago.edu/doi/citedby/10.1086/650205`.

Vincenzo Crupi and Katya Tentori. Confirmation as partial entailment: A representation theorem in inductive logic. *Journal of Applied Logic*, 11(4):364 – 372, 2013. ISSN 1570-8683. doi: http://dx.doi.org/10.1016/j.jal.2013.03.002. URL `http://www.sciencedirect.com/science/article/pii/S1570868313000128`. Combining Probability and Logic: Papers from Progic 2011.

Vincenzo Crupi and Katya Tentori. State of the field: Measuring information and confirmation. *Studies in History and Philosophy of Science Part A*, 47:81–90, 2014.

Vincenzo Crupi, Katya Tentori, and Michel Gonzalez. On bayesian measures of evidential support: Theoretical and empirical issues*. *Philosophy of Science*, 74 (2):229–252, 2007. ISSN 00318248, 1539767X. URL `http://www.jstor.org/stable/10.1086/520779`.

Pierre Maurice Marie Duhem. *The aim and structure of physical theory*. Princeton University Press, Princeton, 1954. ISBN 978-0-691-02524-7.

Richard Feldman. Respecting the evidence. *Philosophical Perspectives*, 19(1):95–119, 2005.

Richard Feldman and Earl Conee. Evidentialism. *Philosophical Studies*, 48(1):15–34, 1985. doi: 10.1007/bf00372404.

Roberto Festa. Bayesian confirmation. In M. C. Galavotti and A. Pagnini, editors, *Experience, Reality, and Scientific Explanation*, pages 55–87. Kluwer Academic Publishers, 1999.

Paul Feyerabend. *Wider den Methodenzwang*. Suhrkamp Taschenbuch Wissenschaft. Suhrkamp, Frankfurt am Main, 1976.

Hartry Field. Recent debates about the a priori. In Tamar Szabo Gendler and John Hawthorne, editors, *Oxford Studies in Epistemology Volume 1*. Oxford University Press, 2005.

Branden Fitelson. Evidence of evidence is not (necessarily) evidence. *Analysis*, 72 (1):85–88, 2012. doi: 10.1093/analys/anr126.

Ronald N. Giere. *Explaining Science*. University of Chicago Press, 2010.

Alvin Goldman and Bob Beddor. Reliabilist Epistemology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

Alvin I Goldman. *Knowledge in a social world*. Clarendon Press ; Oxford University Press, Oxford; New York, 2003. ISBN 978-0-19-823777-8 978-0-19-823820-1.

Patrick Grim and Daniel Singer. Computational Philosophy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition, 2020.

Rainer Hegselmann and Ulrich Krause. Opinion Dynamics and Bounded Confidence Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

Carl G. Hempel. Studies in the logic of confirmation (i.). *Mind*, 54(213):1–26, 1945a.

Carl G. Hempel. Studies in the logic of confirmation (ii.). *Mind*, 54(214):97–121, 1945b.

Franz Huber. Hempel 's logic of confirmation. *Philosophical Studies*, 139:181–189, 2008.

F James. *Statistical methods in experimental physics*. World Scientific, Hackensack, NJ, 2006. ISBN 978-981-256-795-6 978-981-270-527-3. OCLC: 85207490.

Thomas Kelly. Evidence. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2016 edition, 2016.

John G. Kemeny and Paul Oppenheim. Degree of factual support. *Philosophy of Science*, 19(4):307–324, 1952.

Philip Kitcher. *The Advancement of Science : Science without Legend, Objectivity without Illusions*. Oxford University Press, USA, 1993. ISBN 9780198021506.

Philip Kitcher. Patterns of scientific controversies. In Peter K. Machamer, Marcello Pera, and Aristeidēs Baltas, editors, *Scientific Controversies: Philosophical and Historical Perspectives*, page 21. Oxford University Press, 2000.

KIT's Department of Philosophy, ITZ and ITAS. DebateLab - Research Group for the Normative Study of Reasoning. `https://debatelab.philosophie.kit.edu/`, 2017. Online; accessed 23 June 2021.

Thomas S. Kuhn. *Die Struktur wissenschaftlicher Revolutionen*. Number 25 in Suhrkamp-Taschenbuch Wissenschaft. Suhrkamp, Frankfurt am Main, 2., rev. u. um d. postskriptum von 1969 erg. aufl., 6. aufl edition, 1983. ISBN 978-3-518-27625-9.

Imre Lakatos. Falsification and the Methodology of Scientific Research Programmes. In Imre Lakatos and Alan Musgrave, editors, *Criticism and the Growth of Knowledge*, pages 91–195. Cambridge University Press, 1970.

Maria Lasonen-Aarnio. Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, 88(2):314–345, 2014. doi: 10.1111/phpr.12090.

L. Laudan. *Progress and its Problems: Toward a Theory of Scientific Growth*. University of California Press, 1977.

Larry Laudan. *Science and values the aims of science and their role in scientific debate*. University of California Press, Berkeley, 1984. ISBN 978-0-520-05267-3 978-0-520-90811-6. URL `http://site.ebrary.com/id/10676174`.

Larry Laudan. Demystifying underdetermination. In C. Wade Savage, editor, *Scientific Theories*, pages 267–97. University of Minnesota Press, 1990.

John L. Pollock. Reliability and justified belief. *Canadian Journal of Philosophy*, 14(1):103–114, 1984. ISSN 00455091. URL `http://www.jstor.org/stable/40231356`.

Karl Popper. *The Logic of Scientific Discovery*. Routledge, 1935.

Karl R Popper. *Truth, rationality and the growth of scientific knowledge*. Klostermann, Frankfurt am Main, 1979. ISBN 978-3-465-01355-6.

W. V. Quine. Epistemology naturalized. In *Ontological Relativity and Other Essays*. New York: Columbia University Press, 1969.

M. J. S Rudwick. *The great Devonian controversy the shaping of scientific knowledge among gentlemanly specialists*. University of Chicago Press, Chicago, 1988. ISBN 978-0-226-73100-1.

Gerhard Schurz. *Einführung in die Wissenschaftstheorie*. Wissenschaftliche Buchgesellschaft, Darmstadt, 2006. ISBN 978-3-534-15462-3.

Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT). BwUniCluster 2.0. `https://wiki.bwhpc.de/e/Category:BwUniCluster_2.0/`, 2017. Online; accessed 23 June 2021.

William Talbott. Bayesian Epistemology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016a. URL `http://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian/`.

William J. Talbott. A non-probabilist principle of higher-order reasoning. *Synthese*, 193(10):3099–3145, October 2016b. ISSN 0039-7857, 1573-0964. doi: 10.1007/s11229-015-0922-y. URL `http://link.springer.com/10.1007/s11229-015-0922-y`.

Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, 2000.

Eric Winsberg. Computer Simulations in Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2019 edition, 2019.

Wolfgang Stegmüller. *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band 2. Theorie und Erfahrung*. Springer Verlag, 1969.

Wolfram Research, Inc. Mathematica. `https://www.wolfram.com/`, 2019. Online; accessed 23 June 2021.