Logical Reasoning in Social Settings

EDIZIONI
DELLA
NORMALE

67

SEMINARI
E CONVEGNI

Piero Avitabile
IMT School for Advanced Studies Lucca
Piazza S. Francesco 19
55100 Lucca
piero.avitabile@imtlucca.it

Matteo Bizzarri
Scuola Normale Superiore
Piazza dei Cavalieri 7
56126 Pisa, Italia
matteo@bizzarri@sns.it

Lorenzo Casini
Dipartimento di Filosofia
Università di Bologna
Via Zamboni 38
40126 Bologna
lorenzo.casini3@unibo.it

Gustavo Cevolani
IMT School for Advanced Studies Lucca
Piazza S. Francesco 19
55100 Lucca
gustavo.cevolani@imtlucca.it

Davide Chiarella
Istituto di Linguistica Computazionale
Antonio Zampolli
Consiglio Nazionale delle Ricerche
Via De Marini 6
16149 Genova
davide.chiarella@ilc.cnr.it

Marcello D'Agostino
Dipartimento di Filosofia Piero Martinetti
Università degli Studi di Milano
Via Festa del Perdono 7
20122 Milano
marcello.dagostino@unimi.it

Jürgen Landes
Dipartimento di Filosofia Piero Martinetti
Università degli Studi di Milano
Via Festa del Perdono 7
20122 Milano
jurgen.landes@unimi.it

Costanza Larese
Dipartimento di Filosofia Piero Martinetti
Università degli Studi di Milano
Via Festa del Perdono 7
20122 Milano
costanza.larese@unimi.it

Mario Piazza
Scuola Normale Superiore
Piazza dei Cavalieri 7
56126 Pisa, Italia
mario.piazza@sns.it

Carlo Proietti
Istituto di Linguistica Computazionale
Antonio Zampolli
Consiglio Nazionale delle Ricerche
Via De Marini 6
16149 Genova
carlo.proietti@ilc.cnr.it

Andrea Sabatini
Scuola Normale Superiore
Piazza dei Cavalieri 7
56126 Pisa, Italia
andrea.sabatini@sns.it

Caterina Sisti
Scuola Normale Superiore
Piazza dei Cavalieri 7
56126 Pisa, Italiav
caterina.sisti@sns.it

Alejandro Solares-Rojas
Institute for Research in Computer Science
University of Buenos Aires
Pabellón Cero+Infinito, Ciudad Universitaria
Ciudad Autónoma de Buenos Aires, Argentina
asrojas@dc.uba.ar

Matteo Tesi,
Institute of Logic and Computation,
Vienna University of Technology,
Favoritenstraße 9/11, Vienna, Austria
matteo.tesi@tuwien.ac.at

Pietro Vigiani
Scuola Normale Superiore
Piazza dei Cavalieri 7
56126 Pisa, Italia
pietro.vigiani@sns.it

Sofia Elisabetta Walters
Scuola Normale Superiore
Piazza dei Cavalieri 7
56126 Pisa, Italia
sofia.walters@sns.it

Logical Reasoning in Social Settings

edited by
Mario Piazza
Matteo Tesi
Pietro Vigiani

EDIZIONI
DELLA
NORMALE

Workshop
Reasoning with Imperfect Information
in Social Settings
26-28th October 2023
Scuola Normale Superiore, Pisa

First edition: June 2025
e-Isbn 978-88-7642-806-7
doi https://doi.org/10.2422/978-88-7642-806-7

Contents

Introduction: Logic and formal epistemology in social settings

In this chapter, we provide a concise introduction to the topics of the volume. We do so by individuating the main areas of contemporary research in social epistemology, as well as the formal methods used in each area. Moreover, we present the contributions included in the volume within their respective areas of research, divided according to the specific methodology employed. We emphasise that logic can play a key role in social epistemology, in that it can provide the tools to unify areas of research usually pursued with different methods.

The assessment and understanding of fragmented and vague information has become an increasingly pressing issue in social deliberations. In an era of rapid information exchange, rational agents and decision-makers are frequently tasked with navigating inconsistent, incomplete, and sometimes contradictory data, all while striving to organise it into a coherent framework. In such contexts, shared information during communication requires agents to continually update their beliefs, incorporating new evidence while discarding outdated or unreliable data. Furthermore, cognitive biases often distort the information processing, polluting the decision-making process and introducing uncertainty in evaluating evidence.

Addressing these challenges also involves resolving disagreements among agents and managing opinion dynamics, which in turn necessitates evaluating the strength and coherence of various arguments presented within a public discussion. These concerns have become central to a wide range of research areas, where they intersect with debates about rationality, decision theory, and communication.

As a result, these issues have been taken up with a variety of methods, reflecting the multidisciplinary nature of the problem. Philosophers, for instance, employ conceptual analysis to understand the nature of belief revision and rational deliberation, while logicians provide formal, mathematical models to analyse the coherence and consistency of arguments. Psychologists contribute by conducting experiments that examine how humans actually process fragmented or ambiguous information, often revealing the influence of biases and heuristics. Finally, computer-aided simulations and algorithmic approaches have become increasingly prominent in understanding how information spreads, how opinions evolve within groups, and how decisions are made in complex, information-rich environments. Together, these diverse approaches aim to shed light on how agents can navigate the complexities of real-world information to make sound, informed decisions.

The motivation behind this volume stems from the belief that logic, broadly understood, can act as a formal platform capable of unifying diverse, independently pursued lines of research. In particular, when it comes to reasoning within social contexts, logical methods offer a powerful toolkit for the formal analysis of complex social phenomena such as information dynamics, disagreement resolution, and collective decision-making. Guided by this conviction, we organized a workshop in Pisa in October 2023, titled Reasoning with Imperfect Information in Social Settings, under the auspices of the project Understanding Public Data: Experts, Decisions, Epistemic Values, conducted collaboratively by the Scuola Normale Superiore, the IMT School for Advanced Studies in Lucca, and the Institute for Advanced Study of Pavia (IUSS).

The workshop aimed to bring together researchers from the fields of logic, formal and social epistemology, and computer science who are investigating the intricacies of reasoning and information exchange in social settings. The present volume includes most of the papers presented at the workshop and seeks to foster dialogue across these different yet interconnected fields, with the aspiration that social epistemology may benefit from the rigorous application of logical methods.

We divided the contributions in four different groups. While the objects of study are strongly connected, the groups differ from one another from a methodological point of view.

Systems: Logical Frameworks

Papers in the first group utilize logical methods, particularly focusing on proof-theoretic approaches. These three contributions stem from a refined analysis of reasoning models in social and uncertain contexts, exemplified by frameworks such as AGM belief revision (*cf.* [1]) and FDE logic (*cf.* [3, 8]). Each paper employs distinct proof-theoretic methods — including tableaux, sequent calculi, and their generalizations (*cf.* [5]) — to explore these models. A unifying feature of these methods is their reliance on analyticity, meaning the calculi satisfy a form of the subformula property, meaning that the information that circulates in formal proofs is already contained in the conclusion (*cf.* [19]). Analyticity is a tool that allows for information to be managed in a purely finitary and syntactic way, without resorting to semantic intuitions.

In Chapter 2, Marcello D'Agostino, Costanza Larese and Alejandro Solares-Rojas bridge the research program od depth–bounded logics with the non–classical logic of First Degree Entailment. Their work comprises a philosophical presentation of the system for 0-depth FDE, which is argued to provide a model of agents reasoning with incomplete and inconsistent databases and whose space of actually possessed (in contrast with potentially available) information throughout inferences is also incomplete. In this way, the authors' framework allows to formalise inferences in FDE which do not rely on the hypotethical assumption of any virtual information about potential sources. Moreover, D'Agostino, Larese and Solares-Rojas provide a hierarchy of logics, specified by the number of virtual information allowed, converging to FDE.

In Chapter 3, Andrea Sabatini explores AGM belief revision through the lens of structural proof theory. He begins by providing a syntactic characterization of maximally consistent subsets of clause sets, laying the groundwork for a detailed, constructive approach to base-generated belief revision. Building on this foundation, Sabatini develops a hybrid hypersequent calculus for base-generated revision. This calculus arranges sequents and antisequents in parallel, facilitating the formalization of contradictory updates in relation to the provability of extra-logical axioms. Finally, Sabatini proves the admissibility of structural rules, demonstrating that these calculi are both sound and (weakly) complete with respect to a modified, weaker form of the preferential logic R.

In Chapter 4, Matteo Bizzarri employs the framework of fractional semantics in order to model — within a proof-theoretic landscape — notions connected to belief revision. The fractional interpretation of classical propositional logic is obtained through a decomposition procedure within a fully analytic sequent calculus enjoying certain structural properties. This interpretation is here enriched by means of proper axioms used to represent beliefs held by an agent. The addition of these pieces of information preserves the structural properties of the base system. Finally, a further extension involving hyperreal numbers is considered so as to distinguish between full beliefs and revisable ones.

Models: Formal Epistemology

The second group of papers presents various approaches to formally modeling key concepts in formal epistemology (*cf.* [6] for an overview). While numerous formal models of epistemic concepts exist (*cf.* [18] for some salient models), the selected papers represent some foundational methods commonly employed in this field – of course, the list is not exhaustive (*cf.* also [11,14,15,20]). These methods encompass argumentation theory (*cf.* [2, 7]), agent-based modeling (*cf.* [13, 22]), and Bayesian epistemology (*cf.* [4, 21]), each offering distinct frameworks for advancing research in formal epistemology. A unifying feature of these methods is that they offer numerical measures for mathematical descriptions of social scenarios, such as interconnected arguments' strength, polarisation effects and evidence confirmation.

In Chapter 5, Piero Avitabile and Gustavo Cevolani apply a Bayesian epistemological framework to the case of expert testimony. The authors argue that well known (refinements of) Bayes theorem provide an adequate model of credence revision policies in light of uncertain information from an expert. In particular, cases of arguments from authority make it necessary for agents' update policies to consider both the degree of certainty an agent has about the experts' testimony and experts' own reliability towards the testimony (*contra* epistemic deference). Avitabile and Cevolani conclude that Jeffrey conditionalisation is suitable to capture laymen's assessment of experts' testimony.

In Chapter 6, Davide Chiarella and Carlo Proietti investigate which conditions can be conducive to the bi-polarization of opinions in groups. The authors use argumentative agent-based models to model the opinion dynamics in agents. Agent-based modelling has been used to study how the communication and opinion change in a group of agents may lead to a consensus or (on the contrary) to the formation of clusters or polarization effects. Adding and explicitly representing arguments in the model allows to study how the exchange of arguments leads to influence over agents and, possibly, to opinion change.

In Chapter 7, Lorenzo Casini and Jürgen Landes focus on the role played by the conflict of interest (CoI) in evidence. Indeed, evidence affected by CoI has an ambiguous impact: on the one hand, it is biased towards confirmatory results; on the other hand, it is also usually associated with a higher quality, mainly because industry funded research employ large datasets. In view of this, the authors provide a bayesian model for assessing the impact of CoI on evidence. Starting from such model, they show that evidence affected by CoI can still have a confirmatory role for a hypothesis. They conclude arguing that we shall not discard evidence affected by CoI, but rather factor that in based on its reliability.

Foundations: Social Epistemology

The third group collects papers which deal with philosophical aspects of deliberation in social settings. The foundations of social epistemology often rely on data from experimental psychology and cognitive sciences (*cf.* [10,12,16]) to build philosophical theories of rationality (*cf.* [9,17,23]). Theories of uncertain reasoning serves the purpose of laying the groundwork for many themes encountered throughout the volume. Such themes include: how to properly assess disagreement among rational but cognitively bounded agents; the role of experts in aggregating beliefs; how to evaluate biases in opinion formation and its dynamics; how logical standards help mediate between independently formed opinions.

In Chapter 8, Caterina Sisti proposes an analysis of three intersubjective attitudes that groups of agents can have towards a conditional. The author analyses such attitudes in the context of a general theory of extended conditionals, where each conditional is regarded as an implicit device for expressing a belief in the universal generalisation of the extended form of a conditional, where the latter encodes the agent's background knowledge. Sisti then explains agreement, disagreement and weak disagreement in terms of the difference in the amount of agents' background information and the difference in how agents experience the information they have.

In Chapter 9, Sofia Elisabetta Walters presents an adaptive interpretation of the confirmation bias within a model of social reasoning which distinguishes between opinion formation and mediation between opinions. The author understands CB as playing a positive role in opinion formation, as it isolates reasoners and allows them to form their opinions independently. Moreover, standard models of (social) reasoning are criticised as exclusively content-oriented. Walters then suggests that confirmation bias is beneficial to the formation of content of opinions, while mediation is beneficial to the aggregation of opinions. Mediation, in turn, is based on opinion's logical structure, exploiting the intuition that rule–following exerts a normative role on the collective decision process.

References

[1] C. E. Alchourrón, P. Gärdenfors and D. Makinson, *On the logic of theory change: Partial meet contraction and revision functions*, J. Symb. Log. **50** (1985), 510–530.

[2] O. Arieli, A. Borg, J. Heyninck and C. Strasser, *Logic-based approaches to formal argumentation*, J. Appl. Logics **8** (2021), 1793–1898.

[3] N. Belnap, *Modern uses of multiple-valued logic*, In: "Modern Uses of Multiple-Valued Logic", G. M. Dunn and G. Epstein (eds.), D. Reidel Publishing Co., 1977.

[4] L. Bovens and S. Hartmann, "Bayesian Epistemology", Oxford University Press, Oxford, 2003.

[5] A. Ciabattoni, R. Ramanayake andH. Wansing, *Hypersequent and display calculi — a unified perspective*, Studia Logica **102** (2014), 1245–1294.

[6] I. Douven and J. N. Schupbach, *Formal epistemology*, In: "The Oxford Handbook of Topics in Philosophy", Oxford Academic, 2014.

[7] P. M. Dung, *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*, Artificial Intelligence **77** (1995), 321–357.

[8] J. M. Dunn, *Intuitive semantics for first-degree entailment and "coupled trees"*, Philosphical Studies **29** (1976), 149–168.

[9] J. Evans and K. E. Stanovich, *Dual-process theories of higher cognition advancing the debate*, Perspectives on Psychological Science **8** (2013), 223–241.

[10] J. S. B. T. Evans (ed.), "Bias in Human Reasoning: Causes and Consequences", Psychology Press, 1990.

[11] J. Y. Halpern, "Reasoning About Uncertainty", MIT Press, 2003.

[12] D. Kahneman and A. Tversky, *On the reality of cognitive illusions*, Psychological Review **103** (1996), 582–591.

[13] D. Klein, J. Marx and K. Fischbach, *Agent-based modeling in social science, history, and philosophy. an introduction*, Historical Social Research / Historische Sozialforschung **43** (2018), 7–27.

[14] S. Konieczny and R. P. Pérez, *Logic based merging*, J. Philos. Logic **40** (2011), 239–270.

[15] C. List and C. Puppe, *Judgment aggregation: A survey*, In: "Handbook of Rational and Social Choice", P. Anand, P. Pattanaik, and C. Puppe (eds.), Oxford University Press, 2009.

[16] H. Markovits and G. Nantel, *The belief-bias effect in the production and evaluation of logical conclusions*, Memory & Cognition **17** (1989), 11–17.

[17] H. Mercier, *The argumentative theory: Predictions and empirical evidence*, Trends in Cognitive Sciences **20** (2016), 689–700.

[18] R. Pettigrew and J. Weisberg (eds.), "The Open Handbook of Formal Epistemology", PhilPapers Foundation, 2019.

[19] F. Poggiolesi, *On the importance of being analytic: the paradigmatic case of the logic of proofs*, Log. Anal. (N.S.) **55** (2012), 443–461.

[20] L. L. Roland R. Yager (ed.), "Classic Works of the Dempster-Shafer Theory of Belief Functions", Springer, Berlin Heidelberg, 2008.

[21] J. N. Schupbach, "Bayesianism and Scientific Reasoning", Elements in the Philosophy of Science, Cambridge University Press, 2022.

[22] D. Šešelja, *Agent-based models of scientific interaction*, Philosophy Compass **17** (2022), e12855.

[23] D. Sperber and H. Mercier (eds.), "The Enigma of Reason", Harvard University Press, Cambridge, MA, USA, 2017.

Mario Piazza
Matteo Tesi
Pietro Vigiani

Towards more realistic models of logical reasoning. A case study in paraconsistent logic

1. *Introduction*

Theories of rationality are notoriously marred by a discrepancy between theoretical assumptions and practical applications. On the one hand, logical systems model logically omniscient agents (able to correctly recognize all consequences of their assumptions according to the system in use) but provide no sensible means to account for the cost of inferring them. In contrast to omniscient agents, a single practical agent (whether human or artificial) using a certain logic cannot be expected to effectively perform all the correct inferences of that logic, but only those that are within the reach of its limited resources. This is because establishing whether a certain conclusion follows from given premises is often beyond human cognitive resources and, in general, is also computationally hard. This is the case, for instance, of many interesting propositional logics, such as Classical Propositional Logic (**CPL**), First-Degree Entailment (**FDE**) [1], the Logic of Paradox (**LP**) [4, 32], and Strong Kleene Logic ($\mathbf{K_3}$) [29], which are all co-NP complete [3, 12, 36]. Thus, despite the time-honoured idea that logic should guide our reasoning, the problem is that logical systems are somehow "too difficult", in the sense that they provide notions of rationality that cannot be satisfied by human reasoners and resource-bounded agents in general, since they do not account for the cost of reasoning.

A promising step towards more realistic logical models is the *depth-bounded approach* presented and discussed in [19]. It defines a research program inspired by the following core problem:

> *Approximation Problem.* Given a logic L, define a hierarchy of approximating logical systems that converge to L in such a way that these approximations satisfy minimal rational requirement and can be sensibly used as formal models of the deductive power of resource-bounded agents.

In this view, L represents a normative model for an ideal agent with unbounded cognitive and computational resources, while the remaining systems in the hierarchy correspond to decreasing degrees of idealization.

Idealized logics are not just dismissed as descriptively inadequate, but still play a crucial role as limiting normative theories to which approximating models should converge. At the same time, robust implementation of the approximation idea are likely to have a significant practical impact in many research areas — from economics, to philosophy, artificial intelligence and cognitive science — wherever there is an urgent need for more realistic models of deduction.

The depth-bounded approach has been already successfully applied to **CPL** (*e.g.*, [15–17, 19]): the resulting Depth-bounded Boolean Logics are an infinite hierarchy of tractable k-depth approximations to **CPL**, which can be naturally related to a hierarchy of realistic, resource-bounded agents. Two are the key conceptual moves underlying the depth-bounded approach to **CPL**. First, the meaning of a logical operator is specified solely in terms of the information that is actually possessed by an agent, *i.e.*, information practically accessible to her and with which she can operate. This kind of information is called *actual*, and the verb "to hold" is used as synonymous with "to actually possess". Second, a measure of the difficulty of logical deduction can be obtained by focusing on the essential use of "virtual information" and on the depth at which this use is required to infer a given conclusion. Here, virtual information is information that is not actually possessed, but must be temporarily assumed in order to reach a certain conclusion (as, for example, in the discharge rules of natural deduction).

Elaborating on previous work [20], in this paper we discuss in more detail the conceptual and philosophical basis of the depth-bounded approach to **FDE**. In particular, we highlight the more realistic informational view underlying this approach, with special attention to:

(i) the reasons why the Dunn-Belnap's semantics of **FDE** is unsuitable for modeling the reasoning power of realistic and bounded agents;
(ii) the distinction between actual and virtual information in the context of databases which might be incomplete and become inconsistent.

The rest of the paper is organized as follows. Section 2 recalls Dunn's relational semantics and Belnap's four-valued semantics for **FDE** by focusing on the motivations and intended applications. Section 3 provides three reasons why the basic notions of Dunn-Belnap's semantics, while being amenable to epistemic reading, cannot be attained in practice. Section 4 offers a more realistic informational version of **FDE**, by introducing stable imprecise values as the basic notions for the 0-depth logic, *i.e.*, the basic element of the hierarchy of approximations. Both a proof-theoretical characterisation and a 5-valued semantics are provided based on [20]. However, a contribution of this paper consists in improving, simplifying and partially correcting the semantics presented in [20]. Section 5 focuses on the notion of virtual information and discusses k-depth approximations to **FDE**. Final remarks conclude the paper in Section 6.

2. *The semantics of FDE*

The logic of First Degree Entailment (**FDE**) is the core of a family of relevance logics. It was first introduced in the late 1950s by Belnap in his unpublished doctoral thesis and then presented in [1] as a fragment of the system **E** of entailment, namely as a set of relevant implications of the kind $A \to B$, where A and B are implication-free formulae. Several related semantics have been proposed for **FDE** (see [31]). The best-known are Dunn's relational semantics and Belnap's four-valued semantics. The former was introduced by Dunn in the 1960s, but published only later in [23] (by which time it had been proposed also by other authors).

The intuition underlying Dunn's relational semantics is that contradictory propositions should be distinguished from one another:

> [...] let p be the sentence 'It is raining' and let q be the sentence '2 + 2 = 4'. By standard truth table considerations it follows that p & $\sim p$ is true iff p is true and $\sim p$ is true, that is, iff p is true and p is false. Similarly, q & $\sim q$ is true iff q is true and q is false. The question bluntly then is whether the condition that p is true and p is false is the same condition as that q is true and q is false. I think it is not. [...] Intuitively, p & $\sim p$ and q & $\sim q$ describe different situations, granted that neither situation is realizable. [23, p. 154]

With this aim, a Dunn-assignment is defined as a relation η (rather than a function) between propositional variables and the two Boolean values {true, false}. Dunn-assignments are then extended to Dunn-valuations, which are defined as relations between formulae and the two values true and false such that the following conditions are satisfied:

$$\neg A\,\eta\,\text{true iff } A\,\eta\,\text{false} \tag{2.1}$$

$$\neg A\,\eta\,\text{false iff } A\,\eta\,\text{true} \tag{2.2}$$

$$A \wedge B\,\eta\,\text{true iff } A\,\eta\,\text{true and } B\,\eta\,\text{true} \tag{2.3}$$

$$A \wedge B\,\eta\,\text{false iff } A\,\eta\,\text{false or } B\,\eta\,\text{false} \tag{2.4}$$

$$A \vee B\,\eta\,\text{true iff } A\,\eta\,\text{true or } B\,\eta\,\text{true} \tag{2.5}$$

$$A \vee B\,\eta\,\text{false iff } A\,\eta\,\text{false and } B\,\eta\,\text{false} \tag{2.6}$$

A result of this definition is that a formula may relate to true, it may relate to false, it may relate to both, or it may relate to neither. In other words, both gaps and gluts are allowed. Note that the clauses above are exactly the same as the classical truth conditions, without the assumption that truth and falsity are exclusive and exhaustive. Then, the notion of consequence relation is defined as follows: a formula A is a Dunn-consequence of Γ iff for every Dunn-valuation η, if $B\,\eta$ true for all $B \in \Gamma$, then $A\,\eta$ true.

According to Dunn, this framework complies with the intuition discussed above about contradictory propositions, because the two different (of course, unrealizable) situations described by $p \wedge \neg p$ and $q \wedge \neg q$ can be distinguished by a Dunn-valuation η which relates only one of $p \wedge \neg p$ and $q \wedge \neg q$ to *true*. This gives a semantical explanation of one of the principal features of entailment, namely, that $p \wedge \neg p$ need not entail $q \wedge \neg q$ (or an arbitrary formula, say r), for there is a valuation in which $p \wedge \neg p$ is related to *true* and $q \wedge \neg q$ (resp. r) is not. Similarly for tautologies: "rather than thinking about the (*per impossible*) truth conditions for contradictions, we think about the (*per impossible*) 'non-truth' conditions for tautologies" [23, p. 157]. The two different situations described by $p \vee \neg p$ and $q \vee \neg q$ can be distinguished by a Dunn-valuation, η, which does not relate one of $p \vee \neg p$ and $q \vee \neg q$ to *true*. It then follows that $p \vee \neg p$ (or an arbitrary formula r) need not entail $q \vee \neg q$, for there is a valuation in which $p \vee \neg p$ (resp. r) is related to *true*, but $q \vee \neg q$ is not.

Drawing upon the work of Dunn [23] and an observation by Smiley (in correspondence), Belnap put forward a four-valued semantics for **FDE** in [7, 8]. In these papers, he pointed out the usefulness of the resulting characterisation as the logic in which "a computer *should* think". The application that Belnap has in mind is described as follows:

> The reasoner who is to use this logic is an artificial information processor; that is, a (programmed) computer. [...] [2.] The computer is to be some kind of sophisticated question-answering system [that] can also answer questions on the basis of *deductions* which makes from its explicit information. [...] [3.] there is no single, monolithic, infallible source of the computer's data, but the inputs come from several independent sources. [...] [4.] my computer is *not* a complete reasoner, who [...] should, presumably, have some strategy for *giving up* part of what it believes when it finds its belief inconsistent. [...] [5.] In answering its questions, the computer is to reply strictly in terms of what it has been told, *not* in terms of what it could be programmed to believe. [7, pp. 30-32]

For a matrix to characterize a logic adequate for making deductions with information that might be both inconsistent and partial, at least four different values are needed (see [2]). The set of truth-values is $\{\mathbf{t}, \mathbf{f}, \mathbf{b}, \mathbf{n}\}$ is denoted by **4**. These values are interpreted as four possible ways in which an atom p can belong to the present state of information of a computer's database, which in turn is fed by a set Ω of equally "reliable" sources:

- **t** means that the computer is told that p is true by some source, without being told that p is false by any source;

- **f** means that the computer is told that p is false but never told that p is true;
- **b** means that the computer is told that p is true by some source and that p is false by some other source (or by the same source in different times);
- **n** means that the computer is told nothing about the value of p.

In essence, each value represents a subset of the set $\{$true, false$\}$ of the classical values or, equivalently, one of the four possibilities in which a formula A can be related to the values true and false through a Dunn-valuation. These four values form two distinct lattices, depending on whether we consider the partial information ordering induced by set-inclusion (*approximation* lattice) or the partial ordering based on "closeness to the truth" (*logical* lattice). The information ordering is the one according to which the epistemic state of the computer concerning an atom can evolve over time. As Belnap points out:

> When an atomic formula is entered into the computer as either affirmed or denied, the computer modifies its current set-up by adding a "told True" or "told False" according as the formula was affirmed or denied; it does not subtract any information it already has […] In other words, if p is affirmed, it marks p with **t** if p were previously marked with **n**, with **b** if p were previously marked with **f**; and of course leaves things alone if p was already marked either **t** or **b** [7, p. 12].

A *set-up* is simply an assignment to each of the atoms of exactly one of the values in **4**. Using the tables shown in Figure 1, every set-up can be extended to a valuation function $v : F(\mathcal{L}) \to \mathbf{4}$, where $\mathcal{L} = \{\vee, \wedge, \neg\}$, in the usual inductive way. We call this function a **4**-*valuation*. It establishes how the computer is to answer questions about complex formulae based on a set-up. The tables in Figure 1 are obtained by means of considerations related to "Scott's thesis" about approximation lattices [7]. These considerations have an intuitive counterpart. For example, suppose that A is **n** and B is **b**. Then it is not the case that A and B are both true; hence, $A \wedge B$ is not true. But B is false; hence, $A \wedge B$ is false. Thus, $A \wedge B$ is false but not true, therefore **f**. Similarly for the other cases.

$\widetilde{\vee}$	**t**	**f**	**b**	**n**
t	**t**	**t**	**t**	**t**
f	**t**	**f**	**b**	**n**
b	**t**	**b**	**b**	**t**
n	**t**	**n**	**t**	**n**

$\widetilde{\wedge}$	**t**	**f**	**b**	**n**
t	**t**	**f**	**b**	**n**
f	**f**	**f**	**f**	**f**
b	**b**	**f**	**b**	**f**
n	**n**	**f**	**f**	**n**

$\widetilde{\neg}$	
t	**f**
f	**t**
b	**b**
n	**n**

FIGURE 1 **FDE**-tables.

The Belnap-Dunn's matrix $\mathcal{M}_\mathbf{4} =< \mathcal{V}, \mathcal{D}, \mathcal{O} >$ is the matrix for $\mathcal{L}$ where $\mathcal{V} = \mathbf{4}$, $\mathcal{D} = \{\mathbf{t}, \mathbf{b}\}$, and the functions in $\mathcal{O}$ are defined by the tables in Figure 1. (Warning: do not confuse the values **t** and **f** in **4** with true and false. The latter are *local* values referring to the information coming from a source, the former are *global* values, summarizing the epistemic state of the computer with respect to all the sources.) Thus a **4**-*valuation* is a function $v : F(\mathcal{L}) \to \mathbf{4}$ such that for all A, B:

(i) $v(\neg A) = \tilde{\neg}(v(A))$; and

(ii) $v(A \circ B) = \tilde{\circ}(v(A), v(B))$ (where $\circ$ is $\vee$ or $\wedge$). Then, the notion of consequence relation is defined as follows: a formula A is a $\mathcal{M}_\mathbf{4}$-consequence of a set of formulae Γ iff for every **4**-valuation v if $v(B) \in \{\mathbf{t}, \mathbf{b}\}$ for all $B \in \Gamma$, then $v(A) \in \{\mathbf{t}, \mathbf{b}\}$.

3. *Major sources of idealization*

In his paper, Dunn acknowledges that his relational semantics must be given an epistemic interpretation:

> Do not get me wrong – I am not claiming that there are sentences which are in fact both true and false. I am merely pointing out that there are plenty of situations where we suppose, assert, believe, etc., contradictory sentences to be true, and we therefore need a semantics which expresses the truth conditions of contradictions in terms of the truth values that the ingredient sentences would have to take for the contradictions to be true. [23, p. 157]

Similarly, Belnap recognises that the truth-values of his semantics are epistemic in nature:

> My four values are unabashedly epistemic. According to my instructions, sentences are to be marked with either a ***T*** or an ***F***, a ***None*** or a ***Both***, according as to what the computer has been told; or, with only a slight metaphor, according to what it believes or knows. Does this somehow make the enterprise wrong headed? Or not logic? No. Of course these sentences *have* truth-values independently of what the computer has been told; but who can gainsay that the computer cannot *use* the actual truth-values of the sentences in which it is interested? All it can possibly *use* as a basis for inference is what it knows or believes, *i.e.*, what it has been told. [7, p. 47]

However, we claim that, despite its epistemic nature, the Dunn-Belnap's semantics provides a model of reasoning which is highly idealized and thus cannot be used in practice for representing realistic agents characterised by limited information and bounded cognitive and computational resources. There are three sources of major difficulties in Dunn-Belnap's semantics.

The *first* key observation is that, as suggested by Belnap [7, 8], there is no reason to assume that an agent is "told" about the values of *atoms only*.

This is a highly unrealistic restriction. In most practical contexts we may be told that a certain disjunction is true without being told which of the two disjuncts is the true one. Using the words of Dummett:

> I may be entitled to assert "A or B" because I was reliably so informed by someone in a position to know, but if he did not choose to tell me which alternative held good, I could not apply an or-introduction rule to arrive at that conclusion. [...] Hardy may simply not have been able to hear whether Nelson said "Kismet hardy" or "Kiss me Hardy", though he heard him say one or the other: once we have the concept of disjunction, our perceptions themselves may assume an irremediably disjunctive form.

Indeed, this appears to be a distinctive feature of the down-to-earth kind of information with which we practically operate, which has to be distinguished from the context of (constructive) mathematics, where a disjunction $A \vee B$ is provable (*i.e.*, intuitionistically true) if and only if either A is provable or B is provable:

> Unlike mathematical information, empirical information decays at two stages: in the process of acquisition, and in the course of retention and transmission. An attendant directing theatre-goers to different entrances according to the colours of their tickets might even register that a ticket was yellow or green, without registering which it was, if holders of tickets of either colours were to use the same entrance; even our observations are incomplete, in the sense that we do not and cannot take in every detail of what is in our sensory fields. That information decays yet further in memory and in the process of being communicated is evident. In mathematics, any effective procedure remains eternally available to be executed; in the world of our experience, the opportunity for inspection and verification is fleeting [22, pp. 266-278].

Similarly, in most realistic contexts, we can be told that a certain conjunction is false without being told which of the two conjuncts is the false one. As a familiar example, my computer tells me that "either the username or the password is incorrect". To quote Dummett, the machine is "in a position to know" which of the two disjuncts is true, but for security reasons cannot transmit this piece of information to me. In this situation I am a reliable source that can answer positively the query about the disjunction, but I am unable to answer either of the two queries concerning its components.

Second, as we have seen in the previous section, the information ordering of the values in Belnap's semantics is the one according to which the epistemic state of the computer concerning an atom can evolve over time. This means that the values in **4**, except for **b**, cannot be taken as *stable*. An epistemic set-up is just a snapshot of an epistemic state that evolves over time. If we want to consider the truth-values **t**, **f**, **n** as stable we need to assume complete information about the set of sources Ω. Namely, while the meaning of **b** is "*there is* at least a source assenting to p and at least a source dissenting from p" (which is information empirically accessible to x in the sense that x may actually possess this information without a complete knowledge of Ω), the meaning of **t**, **f** and **n** involves information of the kind "*there is no* source such that …", and so requires complete information about the sources in Ω, which may not be empirically accessible to x at any given time.

Thus, despite their epistemic nature, three of the values in **4** are *information-transcendent* when interpreted as timeless. They refer to an "objective" informational situation concerning the domain of all sources, that may well be inaccessible to the computer at any given time. This feature of Belnap's values makes this semantics unsuitable for most practical contexts, where agents might lack a complete knowledge about the sources. For instance, the agent may well be receiving information from an "open" set of sources as they become accessible (even if the information coming from each single source is assumed to be robust). In such a case, the possibility for an agent to come across a source falsifying "there is no source such that …" is always open.

The *third* crucial point is that for the unrestricted language allowing arbitrary formulae involving $\wedge$, $\vee$ and $\neg$, the decision problem for the $\mathcal{M}_{\mathbf{4}}$-consequence relation is co-NP complete [3, 36] and therefore, very likely to be undecidable in practice. This follows from the celebrated result by Cook [12] showing that **CPL** is co-NP complete, together with the fact that the decision problem of inconsistency in **CPL** can be reduced to the decision problem of entailment in **FDE** [20].

4. *0-depth FDE*

In the previous section, we have claimed that **FDE** cannot be used to characterise the reasoning held by realistic agents with limited information and bounded cognitive and computational resources. Moreover, we have provided three reasons why the basic notions of Dunn-Belnap's semantics can be given an epistemic reading, which nonetheless cannot be attained in practice. The notion of *objective information* depicted by the Dunn-Belnap's semantics is unsuitable to vindicate the time-honoured view of logical inference as an information-processing device. What we need instead is a semantics based on the notion of *actual information*, *i.e.*, to put it with Jaakko Hintikka, information that "we actually possess (as distinguished from the information we have available to us) and with which we can in fact operate" [28, p. 229]. Hintikka's idea is that a relation between an agent a and a sentence A obtains if and only if a *actually*, and not only *potentially*, possesses the information that A is true (respectively, false) and can operate with it.

Such a notion of actual information is at the core of the *informational semantics* for **CPL** developed by D'Agostino and co-authors [15–17, 19], in which the meaning of an n-ary logical operator $\star$ is ultimately based on an intuitive, albeit non-deterministic, 3-valued semantics based on purely informational notions.[1] The proof-theoretic characterization given in [15, 16,19] is based on introduction and elimination (intelim) rules that, unlike those of Gentzen-style natural deduction, involve no "discharge" of temporary assumptions. The resulting logical system, which is called *0-depth logic*, consists of the consequence relation associated with the intelim rules only, is computationally easy (tractable).

Can we define in a similar way an *informational* version of **FDE**, namely, a logical system based on the notion of actual information, which, unlike the informational semantics for **CPL**, is suitable to model databases that receive data from multiple sources and have a propensity to become inconsistent? Can we also define suitable tractable approximations to the full **FDE** logic? Inspired by work of D'Agostino [14], Fitting and Avron [5,25,26], we first shift to *signed formulae*, namely, expressions of the form $\mathsf{T}A$, $\mathsf{F}A$, T^*A, F^*A, where A is an unsigned formula. To address the question above, we let signs express *imprecise values* associated with two distinct bipartitions of **4**. Denoting an agent with a and a valuation with v, their intended interpretation is as follows:

- $\mathsf{T}A$ is read as "a holds that A is at least true" and expresses that $v(A) \in \{\mathbf{t}, \mathbf{b}\}$;
- $\mathsf{F}A$ is read as "a holds that A is non-true" and expresses that $v(A) \in \{\mathbf{f}, \mathbf{n}\}$;
- T^*A is read as "a holds that A is non-false" and expresses that $v(A) \in \{\mathbf{t}, \mathbf{n}\}$;
- F^*A is read as "a holds that A is at least false" and expresses that $v(A) \in \{\mathbf{f}, \mathbf{b}\}$.

Crucially, S-formulae of the form $\mathsf{T}A$ or F^*A express information that a may hold even without a complete knowledge of the set sources Ω. These values are "stable" and require no complete information about the set of sources Ω. However, this is not the case of the other two types of S-formulae, T^*A and $\mathsf{F}A$, which involve complete knowledge of Ω and so can only be

[1] This non-deterministic 3-valued semantics was anticipated by Quine [33] and later rediscovered by Crawford and Etherington [13]. However, Quine did not develop this idea into a fully-fledged logical system and did not link it computational complexity. Crawford and Etherington, on the other hand, clearly related their semantics to computational complexity and to tractable approximations of **CPL**, but also failed to provide a convincing systematic development of their intuition.

assumed hypothetically. The choice of such stable imprecise values is implicit in the choice of the set of designated values by Belnap.

Once the shift from precise to imprecise values is completed, we are ready to define a logical system, that we call 0-*depth* ***FDE***, which is based on the notion of actual information, in the sense that it validates only inferences that an agent can draw without making hypotheses about the "objective" informational situation concerning the whole of Ω, *i.e.*, without making hypothetical assumptions that go beyond the information that she holds. The formal details of the 0-depth logic are provided below. Here, we want to anticipate the way in which the 0-depth logic addresses the three problems marring the Dunn-Belnap's semantics discussed in the previous section. *First*, agents can be told not only about the values of atoms, but also about the values of complex sentences. In particular, an agent may hold the information that $\mathsf{T}A \vee B$, but neither the information that $\mathsf{T}A$ nor that $\mathsf{T}B$. Similarly, she may hold the information that $\mathsf{F}^*A \wedge B$, but neither the information that F^*A nor that F^*B. *Second*, the 0-depth logic does not require agents to have complete knowledge about the sources in Ω, but agents are required to draw all the consequences of the information that they actually hold. Here, the meaning of the logical operators is provided in terms of the information that agents actually possess. *Third*, the 0-depth logic turns out to be tractable, that is to say, there is a feasible mechanical procedure to answer to all questions concerning potential 0-depth consequences of a set of assumptions.

4.1. *Intelim sequences*

In what follows we shall use signed formulae (*S-formulae* for short), namely expressions of the form $\mathsf{T}A$, $\mathsf{F}A$, T^*A, F^*A, where A is an unsigned formula, interpreted as specified above. Now, we say that the *conjugate* of $\mathsf{T}A$ is $\mathsf{F}A$ and vice versa, and that the conjugate of T^*A is F^*A and vice versa. Besides, we shall write $\mathsf{T}\Gamma$ for $\{\mathsf{T}A \mid A \in \Gamma\}$. Moreover, we shall use $\varphi, \psi, \theta, \ldots$, possibly with subscripts, as variables ranging over S-formulae; and $X, Y, Z, \ldots$, possibly with subscripts, as variables ranging over sets of S-formulae. Further, let us use $\bar{\varphi}$ to denote the conjugate of φ. Finally, we say that the *unsigned part* of an S-formula is the unsigned formula that results from it by removing its sign. Given an S-formula φ, we denote by φ^u the unsigned part of φ and by X^u the set $\{\varphi^u \mid \varphi \in X\}$.

A natural proof-theoretic characterization of 0-depth logic is obtained by means of the set of introduction and elimination (*intelim*) rules respectively displayed in Table 1 and 2. Then, the analogous 0-depth system for **CPL** in [15,16,19] is characterized by the intelim rules obtained by removing all the starred signs, replacing them with the unstarred signs T and F, interpreted as "only true" and "only false", and eliminating duplicates.

Table 1 Introduction rules for the standard **FDE** connectives.

$\dfrac{\mathsf{F}A}{\mathsf{F}A\wedge B}$	$\dfrac{\mathsf{F}B}{\mathsf{F}A\wedge B}$	$\dfrac{\mathsf{F}^*A}{\mathsf{F}^*A\wedge B}$	$\dfrac{\mathsf{F}^*B}{\mathsf{F}^*A\wedge B}$
$\dfrac{\mathsf{T}A}{\mathsf{T}A\vee B}$	$\dfrac{\mathsf{T}B}{\mathsf{T}A\vee B}$	$\dfrac{\mathsf{T}^*A}{\mathsf{T}^*A\vee B}$	$\dfrac{\mathsf{T}^*B}{\mathsf{T}^*A\vee B}$
$\dfrac{\begin{array}{c}\mathsf{T}A\\ \mathsf{T}B\end{array}}{\mathsf{T}A\wedge B}$	$\dfrac{\begin{array}{c}\mathsf{F}A\\ \mathsf{F}B\end{array}}{\mathsf{F}A\vee B}$	$\dfrac{\begin{array}{c}\mathsf{T}^*A\\ \mathsf{T}^*B\end{array}}{\mathsf{T}^*A\wedge B}$	$\dfrac{\begin{array}{c}\mathsf{F}^*A\\ \mathsf{F}^*B\end{array}}{\mathsf{F}^*A\vee B}$
$\dfrac{\mathsf{T}A}{\mathsf{F}^*\neg A}$	$\dfrac{\mathsf{F}A}{\mathsf{T}^*\neg A}$	$\dfrac{\mathsf{T}^*A}{\mathsf{F}\neg A}$	$\dfrac{\mathsf{F}^*A}{\mathsf{T}\neg A}$

Table 2 Elimination rules for the standard **FDE** connectives.

$\dfrac{\begin{array}{c}\mathsf{F}A\wedge B\\ \mathsf{T}A\end{array}}{\mathsf{F}B}$	$\dfrac{\begin{array}{c}\mathsf{F}A\wedge B\\ \mathsf{T}B\end{array}}{\mathsf{F}A}$	$\dfrac{\begin{array}{c}\mathsf{F}^*A\wedge B\\ \mathsf{T}^*A\end{array}}{\mathsf{F}^*B}$	$\dfrac{\begin{array}{c}\mathsf{F}^*A\wedge B\\ \mathsf{T}^*B\end{array}}{\mathsf{F}^*A}$
$\dfrac{\mathsf{T}A\wedge B}{\mathsf{T}A}$	$\dfrac{\mathsf{T}A\wedge B}{\mathsf{T}B}$	$\dfrac{\mathsf{T}^*A\wedge B}{\mathsf{T}^*A}$	$\dfrac{\mathsf{T}^*A\wedge B}{\mathsf{T}^*B}$
$\dfrac{\begin{array}{c}\mathsf{T}A\vee B\\ \mathsf{F}A\end{array}}{\mathsf{T}B}$	$\dfrac{\begin{array}{c}\mathsf{T}A\vee B\\ \mathsf{F}B\end{array}}{\mathsf{T}A}$	$\dfrac{\begin{array}{c}\mathsf{T}^*A\vee B\\ \mathsf{F}^*A\end{array}}{\mathsf{T}^*B}$	$\dfrac{\begin{array}{c}\mathsf{T}^*A\vee B\\ \mathsf{F}^*B\end{array}}{\mathsf{T}^*A}$
$\dfrac{\mathsf{F}A\vee B}{\mathsf{F}A}$	$\dfrac{\mathsf{F}A\vee B}{\mathsf{F}B}$	$\dfrac{\mathsf{F}^*A\vee B}{\mathsf{F}^*A}$	$\dfrac{\mathsf{F}^*A\vee B}{\mathsf{F}^*B}$
$\dfrac{\mathsf{T}\neg A}{\mathsf{F}^*A}$	$\dfrac{\mathsf{F}\neg A}{\mathsf{T}^*A}$	$\dfrac{\mathsf{T}^*\neg A}{\mathsf{F}A}$	$\dfrac{\mathsf{F}^*\neg A}{\mathsf{T}A}$

First of all, notice that, starting with a set $\mathsf{T}\Gamma$ – or, more generally, from a set of formulae signed with T or F^*, *i.e.* formulae with stable values – there is no way of obtaining S-formulae of the form $\mathsf{F}A$ or T^*A, that is to say, formulae with unstable values. Given that the intelim rules have all a linear format, their application generates *intelim sequences*:

Definition 4.1.

- Given a set X of S-formulae, an *intelim sequence for X* is a sequence $\varphi_1, \ldots, \varphi_m$ of S-formulae such that, for every $i = 1, \ldots, m$, either $\varphi_i \in X$ or it is the conclusion of the application of an intelim rule to preceding S-formulae.
- An intelim sequence is *closed* if it contains an S-formula φ and its conjugate $\bar{\varphi}$; otherwise, it is *open*.

- A *0-depth intelim proof of* φ *from* X is an intelim sequence for X such that φ is the last S-formula in the sequence.
- An *0-depth intelim refutation of* X is a closed intelim sequence for X.

In Figure 2 we show simple examples of intelim sequences, where each assumption is marked with an "@". We shall abuse of the same relation symbol "$\vdash_0$" to denote 0-depth deducibility and refutability, defined as follows:

Definition 4.2. For all X, φ

- φ is *0-depth deducible* from X, $X \vdash_0 \varphi$, iff there is a 0-depth intelim proof of φ from X;
- X is *0-depth refutable*, $X \vdash_0$, iff there is a 0-depth intelim refutation of X.

Several properties of the logic $\vdash_0$ are proved and discussed in [20]. To begin with, the logic $\vdash_0$ is shown to be a (finitary) Tarskian propositional logic; *i.e.*, $\vdash_0$ satisfies reflexivity, monotonicity, cut, and structurality. Moreover, logic $\vdash_0$ is proved to satisfy the subformula property, which is a key property of logical systems in that it allows us to search for proofs or refutations by *analytic methods*; *i.e.*, by considering solely deduction steps involving formulae that are "contained" in the assumptions, or also in the conclusion in the case of proofs. This implies a drastic reduction of the search space which is crucial for the purpose of automated deduction. Last, logic $\vdash_0$ is shown to be tractable, *i.e.*, decidable in practice, and a simple decision procedure for 0-depth refutability and 0-depth deducibility is provided.

$\mathsf{T}\neg(A \vee B)^{@}$	$\mathsf{T}\neg(A \wedge B)^{@}$
$\mathsf{T}\neg C^{@}$	$\mathsf{F}\neg A \vee \neg B^{@}$
$\mathsf{F}^{*} A \vee B$	$\mathsf{F}^{*} A \wedge B$
$\mathsf{F}^{*} A$	$\mathsf{F}\neg A$
$\mathsf{F}^{*} C$	$\mathsf{F}\neg B$
$\mathsf{F}^{*} A \vee C$	$\mathsf{T}^{*} A$
$\mathsf{T}\neg(A \vee C)$	$\mathsf{F}^{*} B$
	$\mathsf{T}^{*} B$
	$\times$

FIGURE 2 Intelim sequences.

4.2. *Non-deterministic semantics*

The signs of our intelim method can be taken as *imprecise* truth-values represented by the pairs $\{\mathbf{t}, \mathbf{b}\}$, $\{\mathbf{t}, \mathbf{n}\}$, $\{\mathbf{f}, \mathbf{b}\}$, $\{\mathbf{f}, \mathbf{n}\}$ — interpreted, respectively, as "at least true", "non-false", "at least false" and "non-true" — that intuitively encode partial information about the standard truth-values in **4** (see [5, 6]). To avoid notational proliferation we use the same symbols $\mathsf{T}, \mathsf{T}^*, \mathsf{F}^*, \mathsf{F}$ for such imprecise values as well as for the signs. We can then reformulate Dunn's semantics in terms of imprecise values as follows:

$$\neg A\,\eta\,\mathsf{T} \text{ iff } A\,\eta\,\mathsf{F}^* \tag{4.1}$$
$$\neg A\,\eta\,\mathsf{T}^* \text{ iff } A\,\eta\,\mathsf{F} \tag{4.2}$$
$$\neg A\,\eta\,\mathsf{F}^* \text{ iff } A\,\eta\,\mathsf{T} \tag{4.3}$$
$$\neg A\,\eta\,\mathsf{F} \text{ iff } A\,\eta\,\mathsf{T}^* \tag{4.4}$$
$$A \vee B\,\eta\,\mathsf{T} \text{ iff } A\,\eta\,\mathsf{T} \text{ or } B\,\eta\,\mathsf{T} \tag{4.5}$$
$$A \vee B\,\eta\,\mathsf{T}^* \text{ iff } A\,\eta\,\mathsf{T}^* \text{ or } B\,\eta\,\mathsf{T}^* \tag{4.6}$$
$$A \vee B\,\eta\,\mathsf{F}^* \text{ iff } A\,\eta\,\mathsf{F}^* \text{ and } B\,\eta\,\mathsf{F}^* \tag{4.7}$$
$$A \vee B\,\eta\,\mathsf{F} \text{ iff } A\,\eta\,\mathsf{F} \text{ and } B\,\eta\,\mathsf{F} \tag{4.8}$$
$$A \wedge B\,\eta\,\mathsf{T} \text{ iff } A\,\eta\,\mathsf{T} \text{ and } B\,\eta\,\mathsf{T} \tag{4.9}$$
$$A \wedge B\,\eta\,\mathsf{T}^* \text{ iff } A\,\eta\,\mathsf{T}^* \text{ and } B\,\eta\,\mathsf{T}^* \tag{4.10}$$
$$A \wedge B\,\eta\,\mathsf{F}^* \text{ iff } A\,\eta\,\mathsf{F}^* \text{ or } B\,\eta\,\mathsf{F}^* \tag{4.11}$$
$$A \wedge B\,\eta\,\mathsf{F} \text{ iff } A\,\eta\,\mathsf{F} \text{ or } B\,\eta\,\mathsf{F} \tag{4.12}$$

An η-valuation is a relation that sasisfies the above conditions as well as the following two "metaconsistency conditions" for all formulae A:

$$\text{It is not the case that } A\,\eta\,\mathsf{T} \text{ and } A\,\eta\,\mathsf{F} \tag{4.13}$$
$$\text{It is not the case that } A\,\eta\,\mathsf{T}^* \text{ and } A\,\eta\,\mathsf{F}^* \tag{4.14}$$

The validity of these conditions can be easily verified by means of the Dunn-Belnap matrix (Figure 1) and it is not difficult to show that this semantics via imprecise values is equivalent to the standard one. Again, the relation η is not a function, for the same formula can be related to more than one value. More specifically, a formula may be related to any pair $\{S_1, S_2\}$ of imprecise values such that $S_1 \cap S_2 \neq \emptyset$. This is equivalent to assigning a precise value to the formula, as in the following table:

if	then
$A\,\eta\,\mathsf{T}$ and $A\,\eta\,\mathsf{T}^*$	A is **t**
$A\,\eta\,\mathsf{T}$ and $A\,\eta\,\mathsf{F}^*$	A is **b**
$A\,\eta\,\mathsf{F}$ and $A\,\eta\,\mathsf{F}^*$	A is **f**
$A\,\eta\,\mathsf{F}$ and $A\,\eta\,\mathsf{T}^*$	A is **n**

But this is not the whole story. In our conceptual framework, the value of a formula A may be completely *undefined* when the agent's information about Ω is insufficient even to establish any of the imprecise values. Suppose, for example, that in absence of complete information, an agent has not yet received any answer to the query "A?" from any of the sources and has no a priori reason to believe that, for any yet unobserved source in Ω, A will never be "told true" or will never be "told false". In this case, none of the partially defined imprecise values can be assigned to A. This may concern any kind of formula, including atomic ones. Moreover, when A is a compound formula, it may well be that both components are assigned imprecise values by η and yet none of the imprecise values can be assigned to it. Suppose for example that $A\,\eta\,\mathsf{F}^*$ and $B\,\eta\,\mathsf{F}$. It is not difficult to check that this assignment is compatible with *any* of the imprecise values for $A \vee B$. It may be that $A \vee B\,\eta\,\mathsf{T}$ in case A is "precisified" into **b**; or it may be that $A \vee B\,\eta\,\mathsf{T}^*$ if B is precisified into **n**; or, else, it may be that $A \vee B\,\eta\,\mathsf{F}^*$ if B is precisified into **f**; finally, it may be that $A \vee B\,\eta\,\mathsf{F}$ if A is precisified into **f**. Under these circumstances we shall say that $A \vee B$ is completely undetermined. In general, we shall write $A\,\eta\,\bot$ to mean that the (possibly atomic) formula A is undetermined. In the context of this paper the symbol "$\bot$" stands for "the undefined value",[2] and should note be confused with **n** which is a defined value.[3]

It is technically convenient to treat $\bot$ as a fifth imprecise value. Intuitively, full indeterminacy or ignorance about a formula amounts to a situation in which *all* imprecise values are admissible. This in the sense that the agent's information is not sufficient to discard any of them and $\bot$ may eventually be precisified into one of them by the development of the agent's querying process. However, while the other imprecise values provide some partial information about the "real" value, $\bot$ is utterly uninformative.

As already mentioned, there is no reason to assume that an agent is "told" about the values of atoms only. In most practical contexts, it may well be that the agent is told by some sources that a certain disjunction is true without being told which of the two disjuncts is the true one; or, analogously, that a certain conjunction is false, without being told which of the two conjuncts is the false one. In such circumstances, the agent is left

[2] This is the intended meaning of $\bot$ in information orders theory, where it typically denotes the bottom element. In logic $\bot$ is often used for the "absurd".

[3] Interestingly, [24] considers a variant of Belnap's 4-valued logic that results from introducing a fifth "undetermined" value corresponding to the case in which the computer is unable, possibly because of a fault, to retrieve the value of a formula. Albeit the motivation is somewhat similar, the proposed solution is quite different and is not related to the problem of defining tractable approximations. Moreover, (i) our explanation of the need for a fifth value is based on epistemic consideration and (ii) we add the fifth imprecise value to the the four imprecise values to characterize tractable subsystems of **FDE** fully in terms of more realistic imprecise values.

in abeyance as far as the values of the components are concerned. Thus, for example, it may well be that an agent is told by some sources that the $A \vee B$ is true, or that $A \wedge B$ is false, without being told anything about the components by *any* of the sources (see Dummett's quotation in Section 3, Quine on the "primitive meaning of the logical operators" [33, pp. 76–77] and p. 7 above). In this case, the components are both undefined, but the disjunction is "at least true" (T) and the conjunction is "at least false" (F^*). This violates conditions (4.5) and (4.7) above. However, like in the username and password example (p. 7), we may assume that all the conditions (4.1)–(4.14) hold, ideally, under *complete* information about all the sources.[4] In other cases, both the components are undefined and so is their disjunction (conjunction). This prompts for the introduction in the η relation of the same kind of non-determinism that is exhibited by the classical operators in an informational setting and cannot be represented by any "iff" conditions (see [19, Chapter 1]).

Let **5** be the set $\{\mathsf{T}, \mathsf{T}^*, \mathsf{F}^*, \mathsf{F}, \bot\}$. Let us call "$\eta$-relatives" of A all the values in **5** that are related to A by η. The tables in Figure 3 display the sufficient conditions for the η-relatives of a compound formula in terms of the η-relatives of its components. In the tables for the binary operators, some entries specify that two signs are *both* η-relatives of the compound formula *conjunctively*. For example when $A \,\eta\, \mathsf{T}$ and $B \,\eta\, \mathsf{T}^*$, $A \vee B$ is related to both T and T^*. As observed above, this amounts to saying that $A \vee B$ is **t**.[5] Other entries present *alternative* values for the compound formula each of which is admissible as a possible η-relative and can be non-deterministically selected depending on the available information. For example, in the table for $\vee$, when $A \,\eta\, \bot$ and $B \,\eta\, \mathsf{F}$, either of $\bot$ and T^* is an admissible η-relative of $A \vee B$. In some informational contexts the disjunction will be undetermined (when B is precisified into **f**), in others it will be T^* (when B is precisified into **n**). Observe that such a precisification of the value of B can attain only under complete knowledge of the reliable sources for B. When both components are totally undetermined, $A \vee B$, may eventually turn out to be T (when either $A \vee B$ is precisified into T) or T^* (when either $A \vee B$ is precisified into T^*), or else remain undetermined. Similar consideration explain the table for $\wedge$, while the table for $\neg$ is self-explanatory. We shall call these tables **5***N-tables*. A *non-deterministic* **5**-*valuation* (**5**N-

[4] We may concede that certain pieces of information, in some contexts, are "irrimediably disjunctive" in nature, in which case (4.5) and (4.7) would be violated even under complete information about the sources. However, we shall not deal with scenario in the present paper.

[5] In a relational context there is no need to introduce precise values, since they can be simulated by relating a formula to two imprecise values. So, in the table for $\vee$, **t** is only an abbreviation of "both T and T^*". Similarly, in the table for $\wedge$, **f** is an abbreviation of "both F and F^*".

$\vee$	T	F	T^*	F^*	$\bot$
T	T	T	$\mathbf{t}$	T	T
F	T	F	T^*	$\bot$	$\bot, \mathsf{T}^*$
T^*	$\mathbf{t}$	T^*	T^*	T^*	T^*
F^*	T	$\bot$	T^*	F^*	$\bot, \mathsf{T}$
$\bot$	T	$\bot, \mathsf{T}^*$	T^*	$\bot, \mathsf{T}$	$\mathsf{T}, \mathsf{T}^*, \bot$

$\wedge$	T	F	T^*	F^*	$\bot$
T	T	F	$\bot$	F^*	$\bot, \mathsf{F}^*$
F	F	F	F	$\mathbf{f}$	F
T^*	$\bot$	F	T^*	F^*	$\bot, \mathsf{F}$
F^*	F^*	$\mathbf{f}$	F^*	F^*	F^*
$\bot$	$\bot, \mathsf{F}^*$	F	$\bot, \mathsf{F}$	F^*	$\mathsf{F}, \mathsf{F}^*, \bot$

$\neg$	
T	F^*
F	T^*
T^*	F
F^*	T
$\bot$	$\bot$

FIGURE 3 5N-tables.

valuation for short) is a relation $\eta : F(\mathcal{L}) \times \mathbf{5}$ that agrees with the **5**N-tables in the following sense:

If $A\,\eta\,x$, then $\neg A\,\eta\,y$ where y is the (only) value corresponding to x in the table for $\neg$.

If $A\,\eta\,x$ and $B\,\eta\,y$ then $A \circ B\,\eta\,z$ for some z displayed in the entry $\langle x, y\rangle$.

Recall that, in our setting, $A \vee B\,\eta\,\mathbf{t}$ is simply an abbreviation for "$A \vee B\,\eta\,\mathsf{T}$ *and* $A \vee B\,\eta\,\mathsf{T}^*$". Similarly, $A \wedge B\,\eta\,\mathbf{f}$ is an abbreviation for "$A \wedge B\,\eta\,\mathsf{F}$ *and* $A \wedge B\,\eta\,\mathsf{F}^*$".

Moreover, a **5**N-valuation must satisfy, in addition to the "metaconsistency conditions" (4.13) and (4.14), also the following additional one:

$$\text{It is not the case that } A\,\eta\,\bot \text{ and } A\,\eta\,x \text{ with } x \in \mathbf{5} \setminus \{\bot\}. \tag{4.15}$$

In other words, no formula can have an informative imprecise value and be completely undetermined at the same time. That a **5**N-valuation η *satisfies* a signed formula SA obviously means that η relates A to the imprecise value corresponding to the sign S. A set X is said to be **5***N-satisfiable* if there is a **5**N-valuation η which satisfies every element of X.

Definition 4.3. For all X, φ,

- φ is a *0-depth consequence* of X, $X \vDash_0 \varphi$, iff for every **5**N-valuation η, η satisfies φ whenever η satisfies all the elements of X;
- X is *0-depth inconsistent*, $X \vDash_0$, iff it is not **5**N-realizable.

Example 4.4. $\{\mathsf{T}(A \vee B) \wedge \neg A\} \nvDash_0 \mathsf{T}B \vee (A \wedge \neg A)$
Any **5**N-valuation η such that

A	B	$A \vee B$	$\neg A$	$(A \vee B) \wedge \neg A$	$A \wedge \neg A$	$B \vee (A \wedge \neg A)$
F^*	$\bot$	T	T	T	F^*	$\bot$

satisfies the premise but not the conclusion. The assignments in columns 3 and 7 are non-deterministic.

Example 4.5. $\{\mathsf{T}\neg(A \wedge B)\} \nvDash_0 \mathsf{T}\neg A \vee \neg B$
Any 5N-valuation such that

A	B	$\neg A$	$\neg B$	$A \wedge B$	$\neg(A \wedge B)$	$\neg A \vee \neg B$
T	$\bot$	F^*	$\bot$	F^*	T	$\bot$

satisfies the premise but not the conclusion. The assignments in columns 5 and 7 are non-deterministic.

It can be shown that:

Proposition 4.6. *$X \vDash_0 \varphi$ if and only if $X \vdash_0 \varphi$.*

The proof can be adapted from the similar proof for the 0-depth approximation of **CPL** [15, 19].

5. *Virtual information: $\boldsymbol{k}$-depth FDE*

As argued above, the intelim rules characterise the 0-depth logic, which validates only inferences that are based on actual information only, without making hypotheses about the objective informational situation concerning the whole set of sources Ω. As a consequence, the intelim rules are all sound, but not complete for full **FDE**.

The full deductive power of **FDE** can be retrieved by allowing agents to use not only their actual information, but also to make hypotheses concerning the whole of Ω. Completeness for full **FDE** is thus obtained by adding only two branching structural rules, according to which we are allowed to:

(i) append both $\mathsf{T}A$ and $\mathsf{F}A$ as sibling nodes to the last element of any intelim sequence;
(ii) append both T^*A and F^*A in a similar way.

These rules are respectively called PB and PB* as they are closely related to a *generalized* Principle of Bivalence. Observe that each application of these rules amounts to making hypotheses about the outcome of a (partial) precisification of an undetermined sentence. So, their unbounded applications amounts to a (partial) precisification of all subformulae involved and yields full **FDE**.

The intuitive meaning of these rules is that one of the two cases must obtain considering the whole of Ω even if the agent has no actual information about which is the case. In this sense, we call the information expressed by each conjugate S-formula *virtual* (as opposed to *actual*) information; *i.e.*, hypothetical information that the agent does not hold, but she temporarily assumes as if she held it.[6]

Intuitively, the more virtual information needs to be invoked, the harder the inference is for the agent, both from the computational and the cognitive viewpoint. Following this idea, D'Agostino and co-authors ([15–17, 19]) defined the *depth* of a valid inference in **CPL** as the number of nested applications of PB. This conceptual step naturally led to the definition of an infinite hierarchy of tractable depth-bounded approximations to **CPL** in terms of the maximum number of nested applications of PB that are allowed. Here, we follow the same strategy by holding that the nested applications of PB and PB* provide a sensible measure of inferential *depth* in **FDE**. Thus, we define an infinite hierarchy of tractable depth-bounded approximations to the logic in terms of the maximum number of nested applications of the branching rules that are allowed.

Such a hierarchy can be intuitively associated with a hierarchy of increasingly idealized agents with more and more — albeit always bounded — cognitive and computational resources or inferential power. Note, however, that the inferential depth associated with an agent is not intended to be interpreted as an upper bound on her inferential power. Rather, it is understood as the maximum depth for which it is guaranteed that, if she possesses the information explicitly carried by the assumptions, she possesses the information explicitly carried by the conclusion.

[6] Notice that for **CPL** only the first rule, PB, with T and F interpreted as "only true" and "only false", makes sense and is sufficient for completeness.

5.1. *Intelim trees*

As anticipated above, we obtain completeness for full **FDE** by adding to the intelim rules the following two branching structural rules, which are respectively called PB and PB* [7]

$$\frac{}{\mathsf{T}A \mid \mathsf{F}A} \qquad \frac{}{\mathsf{T}^*A \mid \mathsf{F}^*A}$$

In what follows, to keep things simple, we require A to be a subformula[8] of $\Gamma \cup \{B\}$, *i.e.* the set consisting of the premises Γ and of the conclusion B of the given inference, and we denote by sub the function that maps any given set Γ of formulae to the set of all its subformulae. However, the "virtual space", namely, the set of formulae that can be introduced through these rules, can be bounded in a variety of ways without loss of completeness (see [19] for various options in the context of **CPL**).

With the addition of PB and PB* to the stock of rules, deductions are represented by downward-growing trees, which brings the method somewhat closer to tableaux. Each application of PB or PB* stands for the introduction of virtual information about the imprecise value of a formula A, which we shall respectively call the PB-*formula* or PB*-*formula*. Note once again that, whereas signed formulae of the form $\mathsf{T}A$ and F^*A represent information that may be empirically obtained (when A turns out to be **b**), signed formulae of the form T^*A and $\mathsf{F}A$ are obtainable only by applying PB or PB*. In turn, any S-formulae appended via those branching rules will be called a *virtual assumption*. Moreover, from our informational viewpoint, the main conceptual advantage of this proof-theoretic characterization consists in that it clearly separates the *intelim rules* that fix the meaning of the connectives in terms of the information that an agent holds from the two *structural rules* that introduce virtual information (PB and PB*). Informally, an *intelim tree based on a set X* of signed formulae is a tree constructed by applying the intelim and the branching rules starting from the S-formulae in X. A branch of an intelim tree is *closed* if it contains an S-formula φ and its conjugate $\bar{\varphi}$; otherwise, it is *open*. An *intelim tree* is said to be *closed* when all its branches are closed; otherwise, it is *open*. An *intelim proof of φ from X* is an intelim tree $\mathcal{T}$ based on X such that φ occurs in all open branches of $\mathcal{T}$. An *intelim refutation of X* is a closed intelim tree $\mathcal{T}$ based on X.

[7] Generalizations of the rule of bivalence have been fruitfully used in the context of many-valued and substructural logics (see [10, 18, 27]).

[8] For every formula A, a *subformula* of A is defined inductively: (i) A is a subformula of A; (ii) for every binary operator $\circ$, if $B \circ C$ is a subformula of A, then so are B and C; (iii) if $\neg B$ is a subformula of A, so is B; (iv) nothing else is a subformula of A.

Note that every refutation of X is, simultaneously, a proof of φ from X, for every φ. This is because there are no open branches and so the condition that φ occurs at the end of all open branches is vacuously satisfied. This is, of course, a kind of explosivity, but it regards *signed* formulae, and it is compatible with the non-explosivity regarding formulae in **FDE**. The reason is that a set consisting only of S-formulae of the form $\mathsf{T}A$ cannot be refuted. First, observe that, starting from a set $\mathsf{T}\Gamma$, there is no way of obtaining S-formulae of the form $\mathsf{F}A$ or T^*A by applying only intelim rules. The only way of obtaining formulae of such forms is by applying PB or PB* and, thus, adding virtual information. Nonetheless, a set $\mathsf{T}\Gamma$ cannot lead to a closed tree even if we add virtual information when unfolding the information contained in $\mathsf{T}\Gamma$. In fact, it can be shown, by induction on the number of nodes that any intelim tree based on a set $\mathsf{T}\Gamma$ has at least one branch containing only S-formulae of the form $\mathsf{T}A$ or F^*A and so it cannot be closed. Only mixed sets, containing S-formulae signed with T and S-formulae signed with F can be refuted. This allows for the analog of proofs *ex-absurdo* of a formula $\mathsf{T}A$ from a set $\mathsf{T}\Gamma$.

The fact that every intelim tree based on a set $\mathsf{T}\Gamma$ is open corresponds to the fact that for every set Γ of unsigned formulae there exists a **4**-valuation such that all the formulae in Γ obtain a designated value.

As in the classical case, the maximum number of nested applications of the branching rules provide a sensible measure of inferential *depth*:

Definition 5.1. The *depth* of an intelim tree $\mathcal{T}$ is the maximum number of virtual assumptions occurring in a branch of $\mathcal{T}$. An intelim tree $\mathcal{T}$ is a *k-depth intelim proof of φ from X* (a *k-depth intelim refutation of X*) if $\mathcal{T}$ is an intelim proof of φ from X (an intelim refutation of X) and $\mathcal{T}$ is of depth k.

Note that a 0-depth intelim tree is nothing but an intelim sequence. Examples of, respectively, two proofs of depth 1, a refutation of depth 1, and a proof of depth 2, are given in Figure 4. Again, each assumption is marked with an "@". The notion of k-depth deducibility depends on the depth at which the use of virtual information is recursively allowed (as well as on the restriction on the virtual space discussed above). So, finally:

Definition 5.2. For all X, φ, for $k > 0$, $X \vdash_k \varphi$ iff $X \cup \{\psi\} \vdash_{k-1} \varphi$ and $X \cup \{\bar{\psi}\} \vdash_{k-1} \varphi$ for some $\psi^u \in \mathsf{sub}(X^u \cup \{\varphi^u\})$.

When $X \vdash_k \varphi$, we say that φ is *deducible at depth k from X*. The above definition covers also the case of k-depth refutability taking $X \vdash_k$ as shorthand for "$X \vdash_k \varphi$ for all φ". When $X \vdash_k$, we say that X *is refutable at depth k*.

We shall abuse of the same relation symbol "$\vdash_k$" to denote k-depth deducibility and refutability.

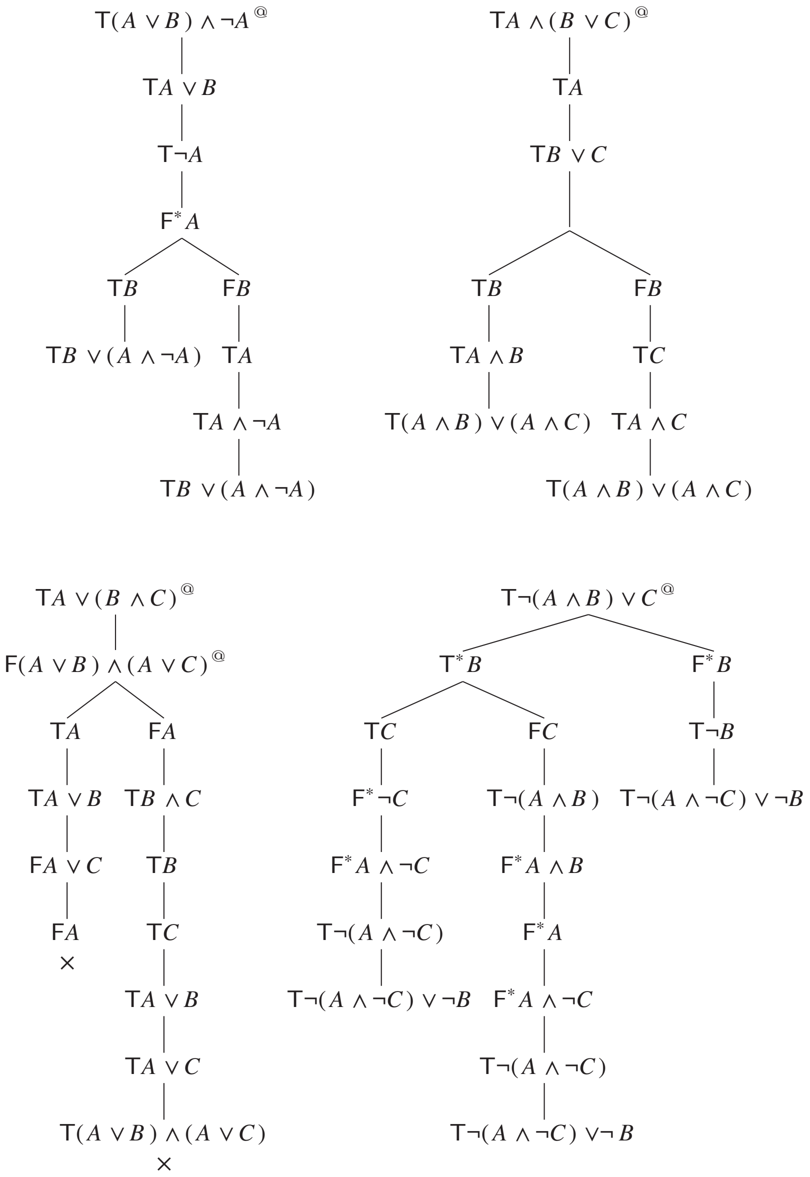

$\mathsf{T}(A \vee B) \wedge \neg A^{@}$
$\mathsf{T}A \vee B$
$\mathsf{T}\neg A$
$\mathsf{F}^{*}A$
$\mathsf{T}B$
$\mathsf{F}B$
$\mathsf{T}B \vee (A \wedge \neg A)$
$\mathsf{T}A$
$\mathsf{T}A \wedge \neg A$
$\mathsf{T}B \vee (A \wedge \neg A)$
$\mathsf{T}A \wedge (B \vee C)^{@}$
$\mathsf{T}A$
$\mathsf{T}B \vee C$
$\mathsf{T}B$
$\mathsf{F}B$
$\mathsf{T}A \wedge B$
$\mathsf{T}C$
$\mathsf{T}(A \wedge B) \vee (A \wedge C)$
$\mathsf{T}A \wedge C$
$\mathsf{T}(A \wedge B) \vee (A \wedge C)$
$\mathsf{T}A \vee (B \wedge C)^{@}$
$\mathsf{F}(A \vee B) \wedge (A \vee C)^{@}$
$\mathsf{T}A$
$\mathsf{F}A$
$\mathsf{T}A \vee B$
$\mathsf{T}B \wedge C$
$\mathsf{F}A \vee C$
$\mathsf{T}B$
$\mathsf{F}A$
×
$\mathsf{T}C$
$\mathsf{T}A \vee B$
$\mathsf{T}A \vee C$
$\mathsf{T}(A \vee B) \wedge (A \vee C)$
×
$\mathsf{T}\neg(A \wedge B) \vee C^{@}$
$\mathsf{T}^{*}B$
$\mathsf{F}^{*}B$
$\mathsf{T}C$
$\mathsf{F}C$
$\mathsf{T}\neg B$
$\mathsf{F}^{*}\neg C$
$\mathsf{T}\neg(A \wedge B)$
$\mathsf{T}\neg(A \wedge \neg C) \vee \neg B$
$\mathsf{F}^{*}A \wedge \neg C$
$\mathsf{F}^{*}A \wedge B$
$\mathsf{T}\neg(A \wedge \neg C)$
$\mathsf{F}^{*}A$
$\mathsf{T}\neg(A \wedge \neg C) \vee \neg B$
$\mathsf{F}^{*}A \wedge \neg C$
$\mathsf{T}\neg(A \wedge \neg C)$
$\mathsf{T}\neg(A \wedge \neg C) \vee \neg B$

Figure 4 k-depth intelim proofs and refutations

Observe that in the above definition the pair of S-formulae, ψ and $\bar{\psi}$, denote a pair of (conjugate) virtual assumptions introduced by respectively PB or PB*. Thus, according to the definition, $X \vdash_k \varphi$ iff the conclusion φ is obtained at depth $k - 1$ by introducing *either* $\mathsf{T}A$ and $\mathsf{F}A$, *or* $\mathsf{T}^{*}A$ and $\mathsf{F}^{*}A$, as virtual assumptions, for some A. This corresponds to the fact that, in our intelim method, a formula φ may be obtained *at a certain depth* by introducing whichever $\mathsf{T}A$ or $\mathsf{F}A$ by an application of PB but not by introducing T^*A or F^*A by an application of PB* and vice versa.

Now, it follows immediately from Definitions 5.1 and 5.2 that for all X, φ, $X \vdash_k \varphi$ ($X \vdash_k$) iff there is a k-depth intelim proof of φ from X (a k-depth intelim refutation of X).

Several properties of the approximations $\vdash_k$ are discussed in [20]. To begin with, the k-depth deducibility relations $\vdash_k$ are shown to satisfy reflexivity, monotonicity, but not cut. However, it is easy to verify that the relations $\vdash_k$ satisfy the following version of cut: if $X \vdash_j \varphi$ and $X \cup \{\varphi\} \vdash_k \psi$, then $X \vdash_{j+k} \psi$. Structurality depends on how the virtual space is restricted, and it is satisfied under our simplifying assumption that PB-formulae and PB*-formulae must belong to $\mathsf{sub}(X^u \cup \{\varphi^u\})$, *i.e.*, must be subformulae of the premises or of the conclusion. Moreover, each approximation $\vdash_k$ satisfies the subformula property and is shown to be tractable. A simple decision procedure for k-depth refutability and k-depth deducibility is also provided.

5.2. *Fixed depth approximations*

Examples 4.4 and 4.5 above are valid inferences in **FDE** that are not so in the 0-depth approximation. Again, the latter is simply the logic of deductive reasoning restricted to the use of actual information. For those valid inferences that cannot be justified solely by the meaning of the connectives — *i.e.*, by the **5**N-tables — the incorporation of virtual information is required. This is information that is not even potentially contained in the current information state. Accordingly, the k-depth logics, $k > 0$, require the simulation of virtual extensions of the current information state.

The k-depth approximation relation is defined by a simple recursion on the basis of the 0-depth consequence relation.

Definition 5.3. For all X, φ, for $k > 0$, $X \vDash_k \varphi$ iff $X \cup \{\psi\} \vDash_{k-1} \varphi$ and $X \cup \{\bar{\psi}\} \vDash_{k-1} \varphi$ for some $\psi^u \in \mathsf{sub}(X^u \cup \{\varphi^u\})$. When $X \vDash_k \varphi$ ($X \vDash_k$), we say that φ is a *k-depth consequence of* X (X *is k-depth inconsistent*).

Given Definition 5.2, it is far from surprising that $X \vdash_k \varphi$ if and only $X \vDash_k \varphi$.

6. *Final remarks*

Approximations to (fragments, full, or extensions of) **CPL** via tractable subsystems of increasing inferential power have been investigated since 1990s (*e.g.* [9, 13, 21, 30, 35]). A hierarchy of tractable depth-bounded approximations to **CPL**, based on an intelim method analogous to the systems presented in this paper, was widely studied in *e.g.*, [15, 16, 19]. This approach has given rise to an articulated research programme at the crossroad of philosophy, mathematical logic and computer science. Its main ideas have been applied and extended in recent works and are at the core of ongoing and future research projects (details can be found in [19]).

This paper, by providing the philosophical motivations and premises to interpret **FDE** in more realistic informational terms, contributes to the project of extending the depth-bounded approach to non-classical contexts. Moreover, as outlined in [20], this work can be naturally extended to the logics **LP** and $\mathbf{K}_3$, which are closely related to **FDE** and can be interpreted along the lines of the informational semantics of **FDE** presented in this paper. Further, the approximations to the simple paraconsistent logics **FDE** and **LP** might serve — both conceptually and technically — as a starting point to extend the approach to other paraconsistent logics such as the logics of formal inconsistency (**LFIs**) [11] and the logics of evidence and truth (**LETs**) [34].

Acknowledgements. This research was funded by the Department of Philosophy "Piero Martinetti" of the University of Milan under the Project "Departments of Excellence 2023–2027" awarded by the Ministry of University and Research (MUR). We also thankfully acknowledge the support of the Italian Ministry of University and Research under the PRIN 2022 program (project n. 2022ZLLR3T). Moreover, Solares-Rojas was funded by CONICET Postdoctoral Fellowships Program, and partly supported by the PAPIIT projects IA105923 and IG400422. Finally, we wish to thank the participants in the discussion at the Pisa workshop for interesting remarks.

References

[1] A. Anderson and N. Belnap, *Tautological entailments*, Philos. Stud. **13** (1962), 9–24.

[2] O. Arieli, A. Avron and A. Zamansky, "Theory of Effective Propositional Paraconsistent Logics", Studies in Logic, Vol. 75, College Publications, London, 2018.

[3] O. Arieli and M. Denecker, *Reducing preferential paraconsistent reasoning to classical entailment*, J. Logic Comput. **13** (2003), 557–580.

[4] F. Asenjo, *A calculus of antinomies*, Notre Dame J. Form. Log. **7** (1966), 103–105.

[5] A. Avron, *Tableaux with four signs as a unified framework*, In: "Automated Reasoning with Analytic Tableaux and Related Methods, Int. Conf., TABLEAUX 2003", M. Cialdea Mayer and F. Pirri (eds.), Lecture Notes in Computer Science, Vol. 2796, Berlin, Heidelberg, Springer, 2003, 4–16.

[6] A. Avron, *Multi-valued semantics: why and how*, Studia Logica **92** (2009), 163–182.

[7] N. Belnap, *How a computer should think*, In: "Contemporary Aspects of Philosophy", G. Ryle (ed.), Oriel Press, Stocksfield, 1977, 30–55.

[8] N. Belnap, *A useful four-valued logic*, In: "Modern Uses of Multiple-Valued Logics", J. Dunn and G. Epstein (eds.), Reidel Publishing Company, Dordrecht, 1977, 5–37.

[9] M. Cadoli and M. Schaerf, *Tractable reasoning via approximation*, Artificial Intelligence **74** (1995), 249–310.

[10] C. Caleiro, J. Marcos and M. Volpe, *Bivalent semantics, generalized compositionality and analytic classic-like tableaux for finite-valued logics*, Texts Theoret. Comput. Sci. EATCS Ser. **603** (2015), 84–110.

[11] W. Carnielli, M. Coniglio and J. Marcos, *Logics of formal inconsistency*, In: "Handbook of Philosophical Logic", D. Gabbay and F. Guenthner (eds.), Springer Netherlands, Dordrecht, 2007, 1–93.

[12] S. Cook, *The complexity of theorem-proving procedures*, In: 'Proceedings of the Third Annual ACM Symposium on Theory of Computing", New York, 1971, 151–158. Association for Computing Machinery.

[13] J. Crawford and D. Etherington, *A non-deterministic semantics for tractable inference*, In" "AAAI/IAAI", AAAI Press / The MIT Press, 1998, 286–291.

[14] M. D'Agostino, "Investigations into the Complexity of some Propositional Calculi", Oxford University, Computing Laboratory, Programming Research Group, 1990.

[15] M. D'Agostino, *An informational view of classical logic*, Theoretical Computer Science **606** (2015), 79–97.

[16] M. D'Agostino, M. Finger and D. Gabbay, *Semantics and proof-theory of depth bounded boolean logics*, Theoret. Comput. Sci. **480** (2013), 43–68.

[17] M. D'Agostino and L. Floridi, *The enduring scandal of deduction*, Synthese **167** (2009), 271–315.

[18] M. D'Agostino, D. Gabbay and K. Broda, *Tableau methods for substructural logics*, In: "Handbook of Tableau Methods", M. D'Agostino, D. Gabbay, R. Hähnle and J. Posegga (eds.), Springer, Dordrecht, 1999, 397–467.

[19] M. D'Agostino, D. Gabbay, C. Larese and S. Modgil, "Depth-bounded Reasoning. Volume 1: Classical Propositional Logic", College Publications, 2024.

[20] M. D'Agostino and A. Solares-Rojas, *Tractable depth-bounded approximations to FDE and its satellites*, J. Logic Comput. (2023).

[21] M. Dalal, *Anytime families of tractable propositional reasoners*, Ann. Math. Artif. Intell. **22** (1998), 297–318.

[22] M. Dummett, “The Logical Basis of Metaphysics”, Duckworth, London, 1991.

[23] J. Dunn, *Intuitive semantics for first-degree entailments and “coupled trees”*, Philos. Stud. **29** (1976), 149–168.

[24] T. M. Ferguson, *Faulty Belnap computers and subsystems of FDE*, J. Logic Comput. **26** (2014), 1617–1636.

[25] M. Fitting, *Bilattices and the semantics of logic programming*, The Journal of Logic Programming **11** (1991), 91–116.

[26] M. Fitting, *Kleene's three valued logics and their children*, Fund. Inform. **20** (1994), 113–131.

[27] R. Hähnle, *Tableaux for many-valued logics*, In: “Handbook of Tableau Methods”, M. D'Agostino, D. Gabbay, R. Hähnle and J. Posegga (eds.), Springer, Dordrecht, 1999, 529–580.

[28] J. Hintikka, “Logic, Language Games and Information: Kantian Themes in the Philosophy of Logic”, Clarendon Press, Oxford, 1973.

[29] S. Kleene, “Introduction to Metamathematics”, Princeton: Van Nostrand, 1952.

[30] G. Lakemeyer and H. Levesque, *A first-order logic of limited belief based on possible worlds*, In: “Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning”, Vol. 17, 2020, 624–635.

[31] H. Omori and H. Wansing, 40 *years of fde: An introductory overview*, Studia Logica **105** (2017), 1021–1049.

[32] G. Priest, *The logic of paradox*, J. Philos. Logic **8** (1979), 219–241.

[33] W. Quine, “The Roots of Reference”, Open Court, LaSalle, 1973.

[34] A. Rodrigues, J. Bueno-Soler and W. Carnielli, *Measuring evidence: a probabilistic approach to an extension of Belnap–Dunn logic*, Synthese **198** (Suppl 22) (2021), 5451–5480.

[35] M. Sheeran and G. Stålmarck, *A tutorial on Stålmarck's proof procedure for propositional logic*, Formal Methods in System Design **16** (2000), 23–58.

[36] A. Urquhart, *The complexity of decision procedures in relevance logic*, In: “Truth or Consequences: Essays in Honor of Nuel Belnap”, J. Dunn and A. Gupta (eds.), Springer, Dordrecht, 1990, 61–76.

Marcello D'Agostino
Costanza Larese
Alejandro Solares-Rojas

Hypersequent calculi for AGM belief revision

1. *Introduction*

Belief change is ubiquitous in ordinary life as well as in scientific theorizing. Historically, a prominent role in the formal analysis of belief change has been played by the framework proposed by Alchourrón, Gärdenfors and Makinson in 1985 [1].

In the AGM model, beliefs are represented by (classically invalid) formulas in a standard formal language for classical propositional logic, and attention is restricted on sets of beliefs closed under classical consequence – typically, under the assumption that a rational agent is committed to believe all the consequences of the beliefs she actually holds [16]. For any belief set $\mathcal{K}$ and any formula A, three types of belief change involving $\mathcal{K}$ and A are taken into consideration:

(i) the *expansion* of $\mathcal{K}$ by A, which results into the smallest logically closed set including $\mathcal{K}$ and A;

(ii) the *contraction* of $\mathcal{K}$ by A, which yields a subset of $\mathcal{K}$ not containing A;

(iii) the *revision* of $\mathcal{K}$ by A, which amounts to the addition of A to $\mathcal{K}$ along with the removal of sentences from $\mathcal{K}$ in order to ensure consistency of the resulting belief set.

In analogy with previous work on propositional default logics [24], this paper introduces a proof-theoretic approach to AGM belief revision centered on a non-standard notion of *hypersequent*.

Hypersequents are lists of sequents separated by a bar, originally conceived to provide analytic calculi for modal and intermediate logics lacking cut-free sequent calculi [2, 4]. We modify the notion of hypersequent in order to embed within derivation trees the consistency checks involved in AGM belief revision: specifically, we redefine hypersequents as *hybrid* constructs, each comprising a sequent and a *set* of *antisequents*. Departing from the conventional disjunctive interpretation of the separating bar, we embrace a conjunctive reading [13, 23]. In this framework, antisequents within a hybrid hypersequent furnish contrary updates concerning the provability of the associated sequent.

The paper is organized as follows. Section 2 introduces the formal apparatus for handling axiomatic extensions of classical logic, namely hybrid sequent calculi with crucial properties. Section 3 contains the proof-theoretic results concerning maximally consistent subsets of sets of clauses that we will subsequently employ. In Section 4, we offer a constructive presentation of base-generated belief revision in terms of maximally consistent subsets of base-generated belief expansion. In Section 5, our focus shifts to hybrid hypersequent calculi. Here, we establish admissibility of structural rules, invertibility of logical rules and the full subformula property for cut-free proofs. We present hybrid hypersequent calculi that are sound and (weakly) complete with respect to (refined) base-generated belief revision, showing that they fail to be strongly complete due to their non-monotonic behaviour in relation to the addition of extra-logical axioms. Finally, in Section 6 we sketch some directions for future research.

2. *Preliminary notions and results*

We use capital Greek letters $\Gamma, \Delta, \Pi, \Sigma, \ldots$ to denote finite *sets* of formulas, and $\Theta, \Lambda \ldots$ to denote sets of *atomic* formulas. For any context Γ we shall be adopting the following conventions: If $\Gamma = \{A_1, A_2, \ldots, A_n\}$, then

$$\Gamma^{\perp} = \{\neg A_1, \neg A_2, \ldots, \neg A_n\} \qquad \bigwedge \Gamma = A_1 \wedge A_2 \wedge \cdots \wedge A_n$$
$$\bigvee \Gamma = A_1 \vee A_2 \vee \cdots \vee A_n.$$

For $\Gamma = \emptyset$, we set $\Gamma^{\perp} = \Gamma$, $\bigwedge \Gamma = \top$, and $\bigvee \Gamma = \perp$, where $\top$ and $\perp$ stand for an arbitrarily chosen tautology and contradiction, respectively. The *logical complexity* $\mathrm{C}(A)$ of a formula A is 1 if A is atomic, $\mathrm{C}(B)+1$ if A is of the form $\neg B$ and $\mathrm{C}(B)+\mathrm{C}(C)+1$ if A is of the form $B \otimes C$, with $\otimes \in \{\wedge, \vee, \rightarrow\}$. The measure C can be easily extended to any multiset $\Gamma = A_1, \ldots, A_n$ by writing $\mathrm{C}(\Gamma) = \mathrm{C}(A_1) + \ldots + \mathrm{C}(A_n)$.

We shall be dealing with Gentzen-style sequents $\Gamma \vdash \Delta$ as well as *antisequents* $\Gamma \dashv \Delta$, where $\Gamma \dashv \Delta$ is valid if, and only if, $\Gamma \vdash \Delta$ is invalid [6, 7, 25]. In the case of classical logic, an antisequent is valid if and only if there exists some Boolean valuation verifying all the formulas in Γ and falsifying all those in Δ.

The system $\overline{\overline{\mathsf{G4}}}$ for classical propositional logic is imported from [21, 27], with logical contexts handled as sets of formulas. In particular, $\overline{\overline{\mathsf{G4}}}$ is obtained by adding to the original Kleene's $\mathsf{G4}$ [14, pp. 289-290, p. 306] the complementary axiom $\dfrac{}{\Theta \dashv \Lambda}\,\overline{ax}$, where $\Theta \cap \Lambda = \varnothing$, as well as distinct rules for antisequents. Whenever generalizing over the union of sequents and antisequents, we write $\Gamma \mathrel{|\!\overset{*}{-}} \Delta$: the measure C can be extended to any (anti)sequent $\Gamma \mathrel{|\!\overset{*}{-}} \Delta$ by writing $\mathrm{C}(\Gamma \mathrel{|\!\overset{*}{-}} \Delta) = \mathrm{C}(\Gamma) + \mathrm{C}(\Delta)$.

AXIOMS

$$\dfrac{}{\Gamma, p \vdash p, \Delta}\,ax \qquad\qquad \dfrac{}{\Theta \dashv \Lambda}\,\overline{ax}$$

LOGICAL RULES

$$\dfrac{\Gamma \vdash \Delta, A}{\Gamma, \neg A \vdash \Delta}\,L\neg \qquad\qquad \dfrac{\Gamma \dashv \Delta, A}{\Gamma, \neg A \dashv \Delta}\,L'\neg$$

$$\dfrac{\Gamma, A \vdash \Delta}{\Gamma \vdash \Delta, \neg A}\,R\neg \qquad\qquad \dfrac{\Gamma, A \dashv \Delta}{\Gamma \dashv \Delta, \neg A}\,R'\neg$$

$$\dfrac{\Gamma, A, B \vdash \Delta}{\Gamma, A \wedge B \vdash \Delta}\,L\wedge \qquad\qquad \dfrac{\Gamma, A, B \dashv \Delta}{\Gamma, A \wedge B \dashv \Delta}\,L'\wedge$$

$$\dfrac{\Gamma \vdash \Delta, A \qquad \Gamma \vdash \Delta, B}{\Gamma \vdash \Delta, A \wedge B}\,R\wedge \qquad\qquad \dfrac{\Gamma \dashv \Delta, A_i}{\Gamma \dashv \Delta, A_1 \wedge A_2}\,R'_i\wedge$$

$$\dfrac{A, \Gamma \vdash \Delta \qquad B, \Gamma \vdash \Delta}{A \vee B, \Gamma \vdash \Delta}\,L\vee \qquad\qquad \dfrac{A_i, \Gamma \dashv \Delta}{A_1 \vee A_2, \Gamma \dashv \Delta}\,L'_i\vee$$

$$\dfrac{\Gamma \vdash \Delta, A, B}{\Gamma \vdash \Delta, A \vee B}\,R\vee \qquad\qquad \dfrac{\Gamma \dashv \Delta, A, B}{\Gamma \dashv \Delta, A \vee B}\,R'\vee$$

$$\dfrac{\Gamma \vdash \Delta, A \qquad \Gamma, B \vdash \Delta}{\Gamma, A \to B \vdash \Delta}\,L\to \qquad \dfrac{\Gamma \dashv \Delta, A}{\Gamma, A \to B \dashv \Delta}\,L'_1\to \qquad \dfrac{B, \Gamma \dashv \Delta}{\Gamma, A \to B \dashv \Delta}\,L'_2\to$$

$$\dfrac{\Gamma, A \vdash \Delta, B}{\Gamma \vdash \Delta, A \to B}\,R\to \qquad\qquad \dfrac{\Gamma, A \dashv \Delta, B}{\Gamma \dashv \Delta, A \to B}\,R'\to$$

FIGURE 1 $\mathsf{G4}$ and $\overline{\overline{\mathsf{G4}}}$ sequent calculi.

Any $\overline{\overline{\mathsf{G4}}}$-derivation π may end either in a sequent $\Gamma \vdash \Delta$ or in an antisequent $\Gamma \dashv \Delta$: in the first case, we say that π is a *proof* for $\Gamma \vdash \Delta$; in the second, π qualifies as a *refutation* for $\Gamma \vdash \Delta$. Furthermore, let us remark that any occurrence of the comma in an (anti)sequent $\Gamma \mathrel{|\!\overset{*}{-}} \Delta$ is interpreted over set-theoretic union: we assume that any potential multiplication of formula occurrences in the conclusion of an application of a $\overline{\overline{\mathsf{G4}}}$ rule is implicitly erased.

Example 2.1. This is a $\overline{\overline{\mathsf{G4}}}$-proof of the sequent $q \vdash ((p \vee \neg p) \vee (p \vee \neg p)) \wedge q$:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{q, p \vdash p}\,ax}{q \vdash p, \neg p}\,\neg_{\mathcal{R}}}{q \vdash p \vee \neg p}\,\vee_{\mathcal{R}}}{q \vdash (p \vee \neg p) \vee (p \vee \neg p)}\,\vee_{\mathcal{R}} \qquad \dfrac{}{q \vdash q}\,ax}{q \vdash ((p \vee \neg p) \vee (p \vee \neg p)) \wedge q}\,\wedge_{\mathcal{R}}$$

We can decompose any (anti)sequent $\Gamma \overset{*}{\vdash} \Delta$ into a set of atomic (anti)sequents, using bottom-up the rules $L\neg, R\neg, L\wedge, R\wedge, L\vee, R\vee, L\to, R\to$ in Figure 1, with $\overset{*}{\vdash}$ in place of $\vdash$, until each leaf of the resulting tree ends with an atomic (anti)sequent.

We recall two crucial features of the $\overline{\overline{\mathsf{G4}}}$ proof system:

Proposition 2.2. $\overline{\overline{\mathsf{G4}}}$ *proves (refutes)* $\Gamma \vdash \Delta$ *if and only if the formula* $\bigwedge \Gamma \to \bigvee \Delta$ *is classically valid (* in*valid).*

Proposition 2.3. *Maximal* $\overline{\overline{\mathsf{G4}}}$*-decomposition yields a unique set of atomic (anti)sequents.*

Proof. For a proof see [3,21]. □

Proposition 2.3 allows us to directly refer to the set of top-clauses associated with a certain (anti)sequent $\Gamma \overset{*}{\vdash} \Delta$, being such a decomposition independent of the specific derivation delivering it. In particular, we write $\mathsf{top}(\Gamma \overset{*}{\vdash} \Delta)$ to indicate the set of top-sequents associated with $\Gamma \overset{*}{\vdash} \Delta$, and $\mathsf{top}_{\mathsf{c}}(\Gamma \overset{*}{\vdash} \Delta)$ to indicate the set of the top-sequents $\Theta \dashv \Lambda \in \mathsf{top}(\Gamma \overset{*}{\vdash} \Delta)$. Moreover, we use $\mathsf{top}^{\star}_{\mathsf{c}}(\Gamma \overset{*}{\vdash} \Delta)$ to denote the complementary clauses in the closure under Cut of $\mathsf{top}(\Gamma \overset{*}{\vdash} \Delta)$, and $cnf(A)$ to refer to the conjunction of the formula translations of the clauses in $\mathsf{top}_{\mathsf{c}}(\overset{*}{\vdash} A)$ (if any) in one-sided format. Finally, if $\mathcal{C}$ is a set of clauses, then we write $\mathcal{C}^{*}$ to denote the closure under Cut of $\mathcal{C}$.

Example 2.4. This is a decomposition-tree of $\overset{*}{\vdash} ((p \to q) \wedge (p \vee s)) \to r$:

$$\dfrac{\overset{*}{\vdash} ((p \to q) \wedge (p \vee s)) \to r}{\dfrac{(p \to q) \wedge (p \vee s) \overset{*}{\vdash} r}{\dfrac{p \to q, p \vee s \overset{*}{\vdash} r}{{}_{L\vee}\dfrac{p \vee s \overset{*}{\vdash} p, r}{p \overset{*}{\vdash} p, r \qquad s \overset{*}{\vdash} p, r} \qquad \dfrac{p \vee s, q \overset{*}{\vdash} r}{p, q \overset{*}{\vdash} r \qquad s, q \overset{*}{\vdash} r}{}_{L\vee}}{}_{L\to}}{}_{L\wedge}}{}_{R\to}$$

Hence, we have that $\mathsf{top}(\overset{*}{\vdash} ((p \to q) \wedge (p \vee s)) \to r) = \{p \overset{*}{\vdash} p, r\, ;\, s \overset{*}{\vdash} p, r\, ;\, p, q \overset{*}{\vdash} r\, ;\, s, q \overset{*}{\vdash} r\}$ with $\mathsf{top}_{\mathsf{c}}(\overset{*}{\vdash} ((p \to q) \wedge (p \vee s)) \to r) = \{s \overset{*}{\vdash} p, r\, ;\, p, q \overset{*}{\vdash} r\, ;\, s, q \overset{*}{\vdash} r\}$ and $cnf(((p \to q) \wedge (p \vee s)) \to r) = (\neg s \vee p \vee r) \wedge (\neg p \vee \neg q \vee r) \wedge (\neg s \vee \neg q \vee r)$.

2.1. *Supraclassical logics*

A (propositional) *supraclassical logic* $\mathcal{S}$ is the extension of (propositional) classical logic with a finite, consistent set of *extra-logical axioms* – i.e., a finite, consistent set of classically invalid formulas. If S is the conjunction of formulas in $\mathcal{S}$ and $\Theta \overset{*}{\vdash} \Lambda$ stands for any clause in $\mathsf{top}^{\star}_{\mathsf{c}}(\overset{*}{\vdash} S)$, then the $\overline{\overline{\mathsf{G4s}}}$ calculus for $\mathcal{S}$ is obtained from $\overline{\overline{\mathsf{G4}}}$ by replacing any instance $\dfrac{}{\Gamma, \Theta \dashv \Lambda, \Delta}$ of the rule $\overline{ax}$ with an instance $\dfrac{}{\Gamma, \Theta \vdash \Lambda, \Delta}$ of the rule ax. Due to Post completeness, each atomic formula occurring in a $\overline{\overline{\mathsf{G4s}}}$-derivation must be interpreted as a propositional constant [19].

Let us say that a rule of the form

$$\frac{\Gamma_1 \vdash \Delta_1 \qquad \cdots \qquad \Gamma_n \vdash \Delta_n}{\Gamma \vdash \Delta}$$

admissible in $\overline{\overline{\mathsf{G4s}}}$ if and only if the sequent $\Gamma \vdash \Delta$ is provable whenever the sequents $\{\Gamma_i \vdash \Delta_i\}_{1 \leq i \leq n}$ are provable.

Proposition 2.5. *The rules of Left and Right Weakening*

$$wk\,\frac{\Gamma \vdash \Delta}{A, \Gamma \vdash \Delta} \qquad\qquad \frac{\Gamma \vdash \Delta}{\Gamma \vdash \Delta, A}\,wk$$

are height-preserving admissible in $\overline{\overline{\mathsf{G4s}}}$.

Proof. By induction on the height of a $\overline{\overline{\mathsf{G4s}}}$-proof π of the premiss (as usual, the height of π is taken to be the number of nodes in a branch of maximal length), exploiting the fact that initial sequents are closed under Weakening. □

Proposition 2.6. *Logical rules of* $\overline{\overline{\mathsf{G4s}}}$ *are height-preserving invertible.*

Proof. By induction on the height of a $\overline{\overline{\mathsf{G4s}}}$-proof π of the premiss. □

Theorem 2.7. *The rule of Cut*

$$\frac{\Gamma \vdash \Delta, A \qquad A, \Pi \vdash \Sigma}{\Pi, \Gamma \vdash \Delta, \Sigma}\,cut$$

is admissible in $\overline{\overline{\mathsf{G4s}}}$.

Proof. We consider the topmost Cut application, and reason by primary induction on the logical complexity of the Cut formula and secondary induction on the sum of the height of the premisses to obtain that both premisses are initial sequents. Notice that proof-transformations must take care of implicit Left and Right Contraction (*cf.* [18]). For instance, take a $\overline{\overline{\mathsf{G4s}}}$-proof π of the form

$$\dfrac{R\wedge\,\dfrac{\overset{\vdots\,\pi_1}{\Gamma \vdash \Delta, A \wedge B, A} \qquad \overset{\vdots\,\pi_2}{\Gamma \vdash \Delta, A \wedge B, B}}{\Gamma \vdash \Delta, A \wedge B} \qquad \dfrac{\overset{\vdots\,\pi_3}{A, B, A \wedge B, \Pi \vdash \Sigma}}{A \wedge B, \Pi \vdash \Sigma}\,L\wedge}{\Pi, \Gamma \vdash \Delta, \Sigma}\,cut$$

We can turn π into the following $\overline{\overline{\mathsf{G4s}}}$-proof π' of $\Pi, \Gamma \vdash \Delta, \Sigma$:

$$\dfrac{\begin{matrix}\vdots_{\pi_4}\\ \Pi, \Gamma \vdash \Delta, \Sigma, B\end{matrix} \quad \dfrac{\text{cut}\,\dfrac{\begin{matrix}\vdots_{\pi_1}\\ \Gamma \vdash \Delta, A \wedge B, A\end{matrix} \quad \wedge_{\mathcal{L}}\dfrac{\begin{matrix}\vdots_{\pi_3}\\ A, B, A \wedge B, \Pi \vdash \Sigma\end{matrix}}{A \wedge B, \Pi \vdash \Sigma}}{\Pi, \Gamma \vdash \Delta, \Sigma, A} \quad \text{cut}\,\dfrac{\dfrac{\begin{matrix}\vdots_{\pi_1}\\ \Gamma \vdash \Delta, A \wedge B, A\end{matrix} \quad \begin{matrix}\vdots_{\pi_2}\\ \Gamma \vdash \Delta, A \wedge B, B\end{matrix}}{\Gamma \vdash \Delta, A \wedge B}\wedge_{\mathcal{R}} \quad \begin{matrix}\vdots_{\pi_3}\\ A, B, A \wedge B, \Pi \vdash \Sigma\end{matrix}}{A, B, \Pi, \Gamma \vdash \Delta, \Sigma}}{B, \Pi, \Gamma \vdash \Delta, \Sigma}\,\text{cut}}{\Pi, \Gamma \vdash \Delta, \Sigma}\,\text{cut}$$

with π_4 being a $\overline{\overline{\mathsf{G4s}}}$-proof of the form

$$\dfrac{\begin{matrix}\vdots_{\pi_2}\\ \Gamma \vdash \Delta, A \wedge B, B\end{matrix} \qquad \dfrac{\begin{matrix}\vdots_{\pi_3}\\ A, B, A \wedge B, \Pi \vdash \Sigma\end{matrix}}{A \wedge B, \Pi \vdash \Sigma}\,L\wedge}{\Pi, \Gamma \vdash \Delta, \Sigma, B}\,\mathit{cut}$$

The Cut application in π_4 is removed by secondary induction, whereas Cut applications on A and B are removed by primary induction.

Subsequently, we prove that the set of all initial sequents is closed under Cut (*cf.* [19, 26]): by Proposition 2.6, it is enough to prove that the set of all initial atomic sequents is closed under Cut.

The most interesting case arises when the the Cut formula is principal in both premisses $\Theta \vdash \Lambda, p$ and $p, \Phi \vdash \Psi$, and $\Phi, \Theta \vdash \Lambda, \Psi$ is not an identity sequent. If $\Theta \mathrel{|\!\!\!\overset{*}{-}} \Lambda, p$ belongs to $\mathsf{top}(\mathrel{|\!\!\!\overset{*}{-}} S)$ and $p, \Phi \mathrel{|\!\!\!\overset{*}{-}} \Psi$ does not belong to $\mathsf{top}(\mathrel{|\!\!\!\overset{*}{-}} S)$, then there exists (at least) one clause $\Phi' \mathrel{|\!\!\!\overset{*}{-}} \Psi'$ in $\mathsf{top}(\mathrel{|\!\!\!\overset{*}{-}} S)$ such that $\Phi' \subseteq \Phi$ and $\Psi' \subseteq \Psi$. If $p \in \Phi'$, then $\Phi' \backslash \{p\}, \Theta \mathrel{|\!\!\!\overset{*}{-}} \Lambda, \Psi'$ belongs to $\mathsf{top}^{\star}_{\mathsf{c}}(\mathrel{|\!\!\!\overset{*}{-}} S)$. If $p \notin \Phi'$, then $\Phi', \Theta \mathrel{|\!\!\!\overset{*}{-}} \Lambda, \Psi'$ is a weakened version of a clause in $\mathsf{top}(\mathrel{|\!\!\!\overset{*}{-}} S)$. On the other hand, if $p, \Phi \mathrel{|\!\!\!\overset{*}{-}} \Psi$ belongs to $\mathsf{top}(\mathrel{|\!\!\!\overset{*}{-}} S)$, then $\Phi, \Theta \mathrel{|\!\!\!\overset{*}{-}} \Lambda, \Psi$ belongs to $\mathsf{top}^{\star}_{\mathsf{c}}(\mathrel{|\!\!\!\overset{*}{-}} S)$ – and we are done. □

Proposition 2.8. $\overline{\overline{\mathsf{G4s}}}$ *proves (refutes) the sequent* $\Gamma \vdash \Delta$ *iff the formula* $\bigwedge \Gamma \to \bigvee \Delta$ *is (is not) a classical consequence of* S.

3. *A proof-theoretic approach to maximally consistent subsets*

In this section, we undertake a syntactic approach to maximally consistent subsets (in short, *mcs*'s) of inconsistent sets of formulas. To improve intelligibility, we introduce *ad hoc* (hybrid) sequent calculi for supraclassical logics featuring explicit structural rules.

Let $\mathcal{S}$ be a supraclassical logic, and $\Theta \mathrel{|\!\!\!\overset{*}{-}} \Lambda$ stand for any clause in $\mathsf{top}_{\mathsf{c}}(\mathrel{|\!\!\!\overset{*}{-}} S)$: the $\overline{\overline{\mathsf{G4str}}}$ calculus for $\mathcal{S}$ is obtained from $\overline{\overline{\mathsf{G4}}}$ by adopting the Weakening and Cut rules as primitive, and replacing any instance $\overline{\Theta \mathrel{|\!\!\!\overset{0}{-}} \Lambda}$ of the rule $\overline{ax}$ with an instance $\overline{\Theta \vdash \Lambda}$ of the rule ax.

Proposition 3.1. $\overline{\overline{\mathsf{G4s}}}$ *proves* $\Gamma \vdash \Delta$ *iff* $\overline{\overline{\mathsf{G4str}}}$ *proves* $\Gamma \vdash \Delta$.

Proof. As for the the right-to-left direction, we reason by induction on the height of a $\overline{\overline{\mathsf{G4str}}}$-proof of $\Gamma \vdash \Delta$, exploiting Propositions 2.5 and Theorem 2.7. □

Proposition 3.2. *Logical rules are invertible in* $\overline{\overline{\mathsf{G4str}}}$.

Proof. By Proposition 3.1 and Proposition 2.6. Notice that preservation of height fails to hold: consider the case of a $\overline{\overline{\mathsf{G4str}}}$-proof π of the form

$$\dfrac{\begin{matrix}\vdots & \qquad & \vdots \\ \Gamma \vdash \Delta, A & \qquad & A, B \to C, \Pi \vdash \Sigma\end{matrix}}{B \to C, \Pi, \Gamma \vdash \Delta, \Sigma}\,\mathit{cut}$$

with $A = C$ and $h(\pi) \leq n + 1$. If we reason by routine induction, there are $\overline{\overline{\mathsf{G4str}}}$-proofs π_1, π_2 of the sequents $A, \Pi \vdash \Sigma, B$ and $A, C, \Pi \vdash \Sigma$, with $h(\pi_1), h(\pi_2) \leq n$: if we apply Cut to the sequents $\Gamma \vdash \Delta, A$ and $A, C, \Pi \vdash \Sigma$ we get a $\overline{\overline{\mathsf{G4str}}}$-proof π' of $\Pi, \Gamma \vdash \Delta, \Sigma$ with $h(\pi') \leq n+1$. If $\overline{\overline{\mathsf{G4}}}$ refutes $\Pi, \Gamma \vdash \Delta, \Sigma$, then we get a $\overline{\overline{\mathsf{G4str}}}$-proof π'' of $C, \Pi, \Gamma \vdash \Delta, \Sigma$ of the form

$$\dfrac{\begin{matrix}\vdots{\scriptstyle \pi'} \\ \Pi, \Gamma \vdash \Delta, \Sigma\end{matrix}}{C, \Pi, \Gamma \vdash \Delta, \Sigma}\,\mathit{wk}$$

with $h(\pi'') \leq n + 2$. □

Lemma 3.3. *If there is a* $\overline{\overline{\mathsf{G4str}}}$*-derivation* π *of* $\Gamma \vdash \Delta$ *from* $\Gamma_1 \vdash \Delta_1, \ldots, \Gamma_n \vdash \Delta_n$*, then* $\overline{\overline{\mathsf{G4str}}}$ *proves* $\bigwedge \Gamma_1 \to \bigvee \Delta_1, \ldots, \bigwedge \Gamma_n \to \bigvee \Delta_n \vdash \bigwedge \Gamma \to \bigvee \Delta$.

Proof. We proceed by induction of $h(\pi)$ to show that there is a $\overline{\overline{\mathsf{G4str}}}$-derivation of $\bigwedge \Gamma_1 \to \bigvee \Delta_1, \ldots, \bigwedge \Gamma_n \to \bigvee \Delta_n, \Gamma \vdash \Delta$ from the sequents $\bigwedge \Gamma_1 \to \bigvee \Delta_1, \ldots, \bigwedge \Gamma_n \to \bigvee \Delta_n, \Gamma_1 \vdash \Delta_1, \ldots, \bigwedge \Gamma_1 \to \bigvee \Delta_1, \ldots, \bigwedge \Gamma_n \to \bigvee \Delta_n, \Gamma_n \vdash \Delta_n$ (the details are omitted). Hence, we infer that there exists a $\overline{\overline{\mathsf{G4str}}}$-derivation of $\bigwedge \Gamma_1 \to \bigvee \Delta_1, \ldots, \bigwedge \Gamma_n \to \bigvee \Delta_n \vdash \bigwedge \Gamma \to \bigvee \Delta$ from $\bigwedge \Gamma_1 \to \bigvee \Delta_1, \ldots, \bigwedge \Gamma_n \to \bigvee \Delta_n, \Gamma_1 \vdash \Delta_1, \ldots, \bigwedge \Gamma_1 \to \bigvee \Delta_1, \ldots, \bigwedge \Gamma_n \to \bigvee \Delta_n, \Gamma_n \vdash \Delta_n$. We exploit Proposition 3.2 to establish that the latter sequents are provable – and this concludes the proof. □

We state that a rule of the form

$$\dfrac{\Gamma_1 \vdash \Delta_1 \qquad \cdots \qquad \Gamma_n \vdash \Delta_n}{\Gamma \vdash \Delta}\,r$$

can be *eliminated* from $\overline{\overline{\mathsf{G4str}}}$ if and only if r belongs to $\overline{\overline{\mathsf{G4str}}}$ and the sequent $\Gamma \vdash \Delta$ is provable without applying r whenever the sequents $\{\Gamma_i \vdash \Delta_i\}_{1 \leq i \leq n}$ are provable without applying r. Furthermore, we say that a Cut application is *inessential* exactly when the Cut formula is not atomic.

Proposition 3.4. *The rule of inessential Cut*

$$\dfrac{\Gamma \vdash \Delta, A \qquad A, \Pi \vdash \Sigma}{\Pi, \Gamma \vdash \Delta, \Sigma}\; cut$$

can be eliminated from $\overline{\overline{\mathsf{G4str}}}$.

Proof. We consider the topmost inessential Cut application, reasoning by induction on the logical complexity of A and exploiting Proposition 3.2. □

Proposition 3.5. *If a clause* $\Theta \vdash \Lambda$ *is provable in* $\overline{\overline{\mathsf{G4str}}}$, *then there is (at least) one* $\overline{\overline{\mathsf{G4str}}}$*-proof* π *of* $\Theta \vdash \Lambda$ *which contains a sequent* $\Phi \vdash \Psi$ *such that*

(i) *every rule applied above* $\Phi \vdash \Psi$ *is* ax *or Cut;*

(ii) *every rule applied below* $\Phi \vdash \Psi$ *is Weakening.*

Proof. Let π' be a $\overline{\overline{\mathsf{G4str}}}$-proof of $\Theta \vdash \Lambda$. If a non-atomic formula A is introduced by the application of a rule in some branch of π', then A occurs down the same branch as an inessential Cut formula – and this contradicts Proposition 3.4: as a result, no logical rule can be applied in π. On the other hand, let us consider the topmost Weakening application in π' which is immediately followed by a Cut application. We reason by cases over the Cut formula p to get the conclusion – *e.g.*, as follows:

$$cut\;\dfrac{wk\;\dfrac{\overset{\vdots}{\Theta' \vdash \Lambda'}}{\Theta' \vdash \Lambda', p} \qquad \overset{\vdots}{p, \Phi' \vdash \Psi'}}{\Phi', \Theta', \vdash \Lambda', \Psi'} \quad \rightsquigarrow \quad \begin{array}{c} \vdots \\ \Theta' \vdash \Lambda' \\ \hline\hline \Phi', \Theta' \vdash \Lambda', \Psi' \end{array}\; wk$$

$$cut\;\dfrac{wk\;\dfrac{\overset{\vdots}{\Theta' \vdash \Lambda', p}}{\Theta' \vdash \Lambda', q, p} \qquad \overset{\vdots}{p, \Phi' \vdash \Psi'}}{\Phi', \Theta', \vdash \Lambda', q, \Psi'} \quad \rightsquigarrow \quad \dfrac{\dfrac{\overset{\vdots}{\Theta' \vdash \Lambda', p} \qquad \overset{\vdots}{p, \Phi' \vdash \Psi'}}{\Phi', \Theta' \vdash \Lambda', \Psi'}\; cut}{\Phi', \Theta' \vdash \Lambda', \Psi', q}\; wk$$

□

Lemma 3.6. *Let $\mathcal{C}$ be a set of clauses. Then $\mathcal{C}$ is inconsistent iff the empty clause belongs to $\mathcal{C}^\star$.*

Proof. For any set $\mathcal{C}$ of clauses, let us say that the $\overline{\overline{\mathsf{G4str}}}$-calculus for $\mathcal{C}$ is the $\overline{\overline{\mathsf{G4str}}}$-calculus for the formula translations of the clauses in $\mathcal{C}$. $(\Rightarrow)$ If $\mathcal{C}$ is inconsistent, Proposition 2.8 ensures that the $\overline{\overline{\mathsf{G4str}}}$ calculus for $\mathcal{C}$ proves the sequents $\vdash A$ and $\vdash \neg A$, for some formula A: by Proposition 3.2 we get the result. □

Theorem 3.7. *Let $\mathcal{C}$ be a set of clauses. Then $\mathcal{D} \subseteq \mathcal{C}$ is a mcs of $\mathcal{C}$ iff for any non-empty clause $p_1, \ldots, p_m \vdash q_1, \ldots, q_n$ in $\mathcal{C} \setminus \mathcal{D}$ the clauses $\{\vdash p_i\}_{1 \leq i \leq m}$ and $\{q_j \vdash\}_{1 \leq j \leq n}$ belong to $\mathcal{D}^\star$.*

Proof. $(\Leftarrow)$ Straightforward from Lemma 3.6. $(\Rightarrow)$ If $\mathcal{D}$ is a mcs of $\mathcal{C}$, then for any non-empty clause $p_1, \ldots, p_m \vdash q_1, \ldots, q_n$ in $\mathcal{C} \setminus \mathcal{D}$ we have that $\mathcal{E} = \mathcal{D} \cup \{p_1, \ldots, p_m \vdash q_1, \ldots, q_n\}$ is an inconsistent set of clauses. If D is the conjunction of the formula translations of the clauses in $\mathcal{D}$, Lemma 3.3 ensures that the $\overline{\overline{\mathsf{G4str}}}$-calculus for $\mathcal{D}$ proves the sequent $D, (p_1 \wedge \cdots \wedge p_m) \to (q_1 \vee \cdots \vee q_n) \vdash \top \to \bot$. Since the $\overline{\overline{\mathsf{G4str}}}$-calculus for $\mathcal{D}$ proves $\vdash D$, it proves also $(p_1 \wedge \cdots \wedge p_m) \to (q_1 \vee \cdots \vee q_n) \vdash \top \to \bot$: we leverage Proposition 3.2 to infer that the $\overline{\overline{\mathsf{G4str}}}$-calculus for $\mathcal{D}$ proves the sequents $\{\vdash p_i\}_{1 \leq i \leq m}$ and $\{q_j \vdash\}_{1 \leq j \leq n}$. Since $\mathcal{D}$ is consistent, Lemma 3.6 ensures that the $\overline{\overline{\mathsf{G4str}}}$-calculus for $\mathcal{D}$ does not prove the empty clause: by Proposition 3.5, we conclude that $\{\vdash p_i\}_{1 \leq i \leq m}$ and $\{q_j \vdash\}_{1 \leq j \leq n}$ belong to $\mathcal{D}^\star$. □

Proposition 3.8. *Let $\mathcal{C}$ be a set of clauses. If $\mathcal{D}$ is a mcs of $\mathcal{C}$, then we have that*

(i) *$(\mathcal{C} \setminus \mathcal{D})$ is a set of complementary clauses;*
(ii) *$(\mathcal{C} \setminus \mathcal{D}) = (\mathcal{C} \setminus \mathcal{D})^*$;*
(iii) *if $\mathcal{C}$ is a set of non-empty clauses, then $\mathcal{C} \setminus \mathcal{D}$ is consistent.*

Proof. (i) If there is an identity clause $\Theta, p \vdash p, \Lambda$ in $\mathcal{C} \setminus \mathcal{D}$, then Theorem 3.7 implies that clauses $\vdash p$ and $p \vdash$ belong to $\mathcal{D}^\star$: as a result, the empty sequent belongs to $\mathcal{D}^\star$ – by Lemma 3.6, a contradiction. (ii) If clauses of the form $\Theta \vdash \Lambda, p$ and $p, \Phi \vdash \Psi$ belong to $\mathcal{C} \setminus \mathcal{D}$, then Theorem 3.7 entails that clauses $\vdash p$ and $p \vdash$ belong to $\mathcal{D}^\star$ – as before, a contradiction. (iii) Suppose by contradiction that $\mathcal{C} \setminus \mathcal{D}$ is inconsistent, and thus that there exists (at least) one non-empty mcs $\mathcal{E}$ of $\mathcal{C} \setminus \mathcal{D}$ such that $\mathcal{E} \subset (\mathcal{C} \setminus \mathcal{D})$. By Theorem 3.7, for any non-empty clause $p_1, \ldots, p_m \vdash$

$q_1, \ldots, q_n$ in $(\mathcal{C} \setminus \mathcal{D}) \setminus \mathcal{E}$ the clauses $\{\vdash p_i\}_{1 \leq i \leq m}$ and $\{q_j \vdash\}_{1 \leq j \leq n}$ belong to $\mathcal{E}^\star$. By Lemma 3.6 and Proposition 3.5, this means that there exist clauses $\{\Theta_i \vdash p_i, \Lambda_i\}_{1 \leq i \leq m}$ and $\{\Phi_j, q_j \vdash \Psi_j\}_{1 \leq j \leq n}$ in $\mathcal{E}$: Theorem 3.7 entails that the clauses $\{p_i \vdash\}_{1 \leq i \leq n}$ and $\{\vdash q_j\}_{1 \leq j \leq n}$ belong to $\mathcal{D}^\star$. On the other hand, by Theorem 3.7 we have that also the clauses $\{\vdash p_i\}_{1 \leq i \leq m}$ and $\{q_j \vdash\}_{1 \leq j \leq n}$ belong to $\mathcal{D}^\star$ – a contradiction. □

Proposition 3.9. *Let $\mathcal{C}$ be a set of clauses. If $\mathcal{C} = \mathcal{C}^\star$ and $\mathcal{D}$ is a mcs of $\mathcal{C}$, then $\mathcal{D} = \mathcal{D}^\star$.*

Proof. Suppose by contradiction that there is (at least) one mcs $\mathcal{D}$ of $\mathcal{C}$ such that $\mathcal{D}^\star \not\subseteq \mathcal{D}$. This implies that there is (at least) one non-empty clause $\Theta \vdash \Lambda$ in $\mathcal{D}^\star$ which belongs to $\mathcal{C} \setminus \mathcal{D}$: by Theorem 3.7, if $\Theta = \{p_1, \ldots, p_m\}$ and $\Lambda = \{q_1, \ldots, q_n\}$, then the clauses $\{\vdash p_i\}_{1 \leq i \leq m}$ and $\{q_j \vdash\}_{1 \leq j \leq n}$ belong to $\mathcal{D}^\star$ – by Lemma 3.6, a contradiction. □

Proposition 3.10. *Let $\mathcal{C}$ be a set of clauses. Then we have that:*

(i) *if $\mathcal{D}$ is a mcs of $\mathcal{C}$, $\mathcal{D}^\star$ may* not *be a mcs of $\mathcal{C}^\star$;*
(ii) *if $\mathcal{D}^\star$ is a mcs of $\mathcal{C}^\star$, $\mathcal{D}$ may* not *be a mcs of $\mathcal{C}$.*

Proof. For each statement we offer an example. (i) Let $\mathcal{C}$ be $\{p, q \vdash r, s;\ r \vdash;\ s \vdash;\ \vdash p;\ \vdash q\}$. By Theorem 3.7, the set $\mathcal{D} = \{\ p, q \vdash r, s;\ r \vdash;\ s \vdash;\ \vdash p\ \}$ is a mcs of $\mathcal{C}$. On the other hand, $\mathcal{D}^\star = \{\ p, q \vdash r, s;\ r \vdash;\ s \vdash;\ \vdash p;\ p, q \vdash;\ p, q \vdash r;\ p, q \vdash s;\ q \vdash r, s;\ q \vdash s;\ q \vdash r;\ q \vdash\ \}$. By Theorem 3.7, $\mathcal{D}^\star$ is *not* a mcs of $\mathcal{C}^\star$, since e.g. $\vdash r$ belongs to $\mathcal{C}^\star \setminus \mathcal{D}^\star$ and $r \vdash$ does not belong to $\mathcal{D}^\star$. (ii) Let $\mathcal{C}$ be $\{\ p \vdash q;\ q \vdash r;\ \vdash p;\ \vdash q;\ q \vdash;\ r \vdash\ \}$, and thus $\mathcal{C}^\star = \{\ p \vdash q;\ q \vdash r;\ \vdash p;\ \vdash q;\ q \vdash;\ r \vdash;\ p \vdash r;\ \vdash r;\ \vdash p;\ \vdash\ \}$. The set $\mathcal{E} = \{\ p \vdash q;\ q \vdash r;\ p \vdash;\ \vdash r;\ q \vdash;\ p \vdash r\ \}$ is a mcs of $\mathcal{C}^\star$: as witnessed by Proposition 3.9, $\mathcal{E} = \mathcal{E}^\star$ – whereas $\mathcal{E}$ is not even a subset of $\mathcal{C}$. □

Proposition 3.11. *If $\mathcal{C}$ is a set of clauses and $\mathcal{D}_1, \ldots, \mathcal{D}_n$ are the mcs's of $\mathcal{C}$, then the following conditions hold:*

(i) *$\mathcal{D}_i \cup \mathcal{D}_j$ is inconsistent, for any $1 \leq i \neq j \leq n$;*
(ii) *if $\mathcal{C} = \mathcal{C}^\star$, then $(\mathcal{D}_1 \cap \ldots \cap \mathcal{D}_i) = (\mathcal{D}_1 \cap \ldots \cap \mathcal{D}_i)^\star$, for any $1 \leq i \leq n$.*

Proof. (i) $\mathcal{D}_i$ and $\mathcal{D}_j$ are distinct mcs's of $\mathcal{C}$: as a result, there exists (at least) one non-empty clause $\Theta \vdash \Lambda$ which belongs to (say) $\mathcal{D}_i$ and not to $\mathcal{D}_j$. The clause $\Theta \vdash \Lambda$ thus belongs to $\mathcal{C} \setminus \mathcal{D}_j$: if $\Theta = p_1, \ldots, p_m$ and $\Lambda = q_1, \ldots, q_{m'}$, then Theorem 3.7 guarantees that the clauses $\{\vdash p_h\}_{1 \leq h \leq m}$ and $\{q_k \vdash\}_{1 \leq k \leq m'}$ belong to $\mathcal{D}_j^\star$. (ii) If $\Theta \vdash \Lambda, p$ and $p, \Phi \vdash \Psi$ belong to $\mathcal{D}_1 \cap \ldots \cap \mathcal{D}_i$, then $\Theta \vdash \Lambda, p$ and $p, \Phi \vdash \Psi$ belong to $\mathcal{D}_j$, for each $1 \leq j \leq i$. By Proposition 3.9 we have that $\mathcal{D}_j = \mathcal{D}_j^\star$: this implies that $\Phi, \Theta \vdash \Lambda, \Psi$ belongs to each $\mathcal{D}_j$ – and we are done. □

4. *Base-generated belief revision*

We say that $\mathcal{B}$ is a *belief base* if $\mathcal{B}$ is a finite, non-empty set of classically invalid formulae, and that $\mathcal{K}$ is a *belief set* if $\mathcal{K}$ is a set of formulae comprising (at least) one extra-logical axiom and which is closed under classical consequence. Furthermore, we state that a belief set $\mathcal{K}$ is *generated by a (belief) base* $\mathcal{B}$ exactly when $\mathcal{K} = Cn(\mathcal{B})$.

Definition 4.1. Let $\mathcal{B}$ be a belief base. The *base expansion of* $\mathcal{B}$ *by* A, in symbols $\mathcal{B} + A$, is $\mathcal{B} \cup \{A\}$.

For any base $\mathcal{B}$ we can axiomatically define the operation of contraction:

Definition 4.2. Let $\mathcal{B}$ be a belief base. The set of formulae $\mathcal{B} \operatorname{div} A$ is the *base contraction of* $\mathcal{B}$ *by* A iff the following postulates are satisfied:

BC1 *Success*: if A is not tautological, then $A \notin (\mathcal{B} \operatorname{div} A)$.
BC2 *Inclusion*: $(\mathcal{B} \operatorname{div} A) \subseteq \mathcal{B}$.
BC3 *Relevance*: if $B \in \mathcal{B}$ and $B \notin (\mathcal{B} \operatorname{div} A)$, then there is a set $\mathcal{B}'$ such that $(\mathcal{B} \operatorname{div} A) \subseteq \mathcal{B}' \subseteq \mathcal{B}$, $A \notin Cn(\mathcal{B}')$ and $A \in Cn(\mathcal{B}' \cup \{B\})$.
BC4 *Uniformity*: if for all $\mathcal{B}' \subseteq \mathcal{B}$ we have that $A \in Cn(\mathcal{B}')$ iff $B \in Cn(\mathcal{B}')$, then $\mathcal{B} \operatorname{div} A = \mathcal{B} \operatorname{div} B$.

Axioms BC1 and BC2 ensure that (at least) A is removed from $\mathcal{B}$ after contraction by A. By Axiom BC3, if a formula B is removed from $\mathcal{B}$ after contraction by A, then B plays some role for the fact that $\mathcal{B}$ logically implies A: in other words, the exclusion of formulas from $\mathcal{B}$ after contraction by A is blocked unless there is some good reason for the exclusion[1]. Finally, Axiom BC4 guarantees that the result of contraction of $\mathcal{B}$ by A depends only on which subsets of $\mathcal{B}$ logically imply A.

On the other hand, one can pursue a constructive approach to base contraction of $\mathcal{B}$ by A *via* the following notions.

Definition 4.3. Let $\mathcal{B}$ be a belief base, and A a formula.

(i) The *remainder set of* $\mathcal{B}$ *with respect to* A, in symbols $\mathcal{B} \bot A$, contains any $\mathcal{B}' \subseteq \mathcal{B}$ for which $A \notin Cn(\mathcal{B}')$ and there is no $\mathcal{B}'' \subseteq \mathcal{B}$ such that $\mathcal{B}' \subset \mathcal{B}''$ and $A \notin Cn(\mathcal{B}'')$.
(ii) A *selection function* for $\mathcal{B}$ is a function γ such that $\gamma(\mathcal{B} \bot A) \neq \varnothing$ and $\gamma(\mathcal{B} \bot A) \subseteq (\mathcal{B} \bot A)$, if $(\mathcal{B} \bot A) \neq \varnothing$ – and $\gamma(\mathcal{B} \bot A) = \mathcal{B}$, otherwise.

Theorem 4.4. *Let* $\mathcal{B}$ *be a belief base. Then* $\mathcal{B} \operatorname{div} A = \bigcap \gamma(\mathcal{B} \bot A)$ *for some selection function for* $\mathcal{B}$.

Proof. For a proof see [10, pp. 657-658]. □

[1] In AGM contraction, the analogous of BC3 is the much-debated axiom of Recovery (*cf.* [9, pp. 219-221] for a brief comparison).

For any base $\mathcal{B}$ we can axiomatically define the operation of revision:

Definition 4.5. Let $\mathcal{B}$ be a belief base. The set of formulae $\mathcal{B} * A$ is the *base revision of* $\mathcal{B}$ *by* A[2] iff the following postulates are satisfied:

BR1 *Success*: if A is not tautological, then $A \in (\mathcal{B} * A)$.
BR2 *Inclusion*: $(\mathcal{B} * A) \subseteq (\mathcal{B} + A)$.
BR3 *Relevance*: if $B \in \mathcal{B}$ and $B \notin (\mathcal{B} * A)$, then there is a consistent set $\mathcal{B}'$ such that $(\mathcal{B} * A) \subseteq \mathcal{B}' \subseteq (\mathcal{B} + A)$ and $\mathcal{B}' + B$ is inconsistent.
BR4 *Uniformity*: if for all $\mathcal{B}' \subseteq \mathcal{B}$ we have that $\neg A \in Cn(\mathcal{B}')$ iff $\neg B \in Cn(\mathcal{B}')$, then $\mathcal{B} * A = \mathcal{B} * B$.

One can undertake a constructive approach to base revision by exploiting the following result:

Theorem 4.6. *If* $\mathcal{B}$ *is a belief base, then* $\mathcal{B} * A = (\mathcal{B} \operatorname{div} \neg A) + A$.

Proof. For a proof see [10, pp. 661-662]. □

Proposition 4.7. *Let* $\mathcal{B}$ *be a belief base and* A *a consistent formula. Then* $\mathcal{B} * A = \bigcap_{i=1}^{n} \mathcal{B}_i$, *for some mcs's* $\mathcal{B}_1, \ldots, \mathcal{B}_n$ *of* $\mathcal{B} + A$ *containing* A.

Proof. Theorem 4.6 and Theorem 4.4 ensure the existence of (at least) one selection function γ for $\mathcal{B}$ such that $\mathcal{B} * A = \left(\bigcap \gamma(\mathcal{B} \bot \neg A)\right) + A$. If $\gamma(\mathcal{B} \bot \neg A) = \{\mathcal{B}'_1, \ldots, \mathcal{B}'_n\}$, then the following equalities hold:

$$\mathcal{B} * A = \left(\bigcap_{i=1}^{n} \mathcal{B}'_i\right) \cup \{A\} = \bigcap_{i=1}^{n} (\mathcal{B}'_i \cup \{A\})$$

Suppose by contradiction that $\mathcal{B}$ contains a formula $B \notin \mathcal{B}'_i$ such that $\mathcal{B}'_i \cup \{A\} \cup \{B\}$ is consistent, for some i. This implies that $\neg A \notin Cn(\mathcal{B}'_i \cup \{B\})$, whereas Definition 4.3 guarantees that $\neg A \in Cn(\mathcal{B}')$, for any $\mathcal{B}' \subseteq \mathcal{B}$ such that $\mathcal{B}'_i \subset \mathcal{B}'$: as a result, we get that $\mathcal{B}'_i \cup \{A\}$ is a mcs of $\mathcal{B} \cup \{A\}$, for any i – as desired. □

In this paper, we shall focus on belief sets generated by belief bases. Base expansion, base contraction and base revision give rise to analogous operations on the generated belief set:

Definition 4.8. Let $\mathcal{B}$ a belief base, and $\mathcal{K} = Cn(\mathcal{B})$. The set of formulae

(i) $\mathcal{K} + A$ is the *base-generated expansion of* $\mathcal{K}$ *by* A iff $\mathcal{K} + A = Cn(\mathcal{B} + A)$;

[2] We use 'base revision of $\mathcal{B}$ by A' to denote what in the literature is known as '*internal* (partial meet) base revision of $\mathcal{B}$ by A' (cf. [10, p. 649]).

(ii) $\mathcal{K} \operatorname{div} A$ is the *base-generated contraction of $\mathcal{K}$ by A* iff $\mathcal{K} \operatorname{div} A = Cn(\mathcal{B} \operatorname{div} A)$;
(iii) $\mathcal{K} * A$ is the *base-generated revision of $\mathcal{K}$ by A* iff $\mathcal{K} * A = Cn(\mathcal{B} * A)$.

Base-generated contraction of a belief set can be axiomatically characterized as follows:

Theorem 4.9. *Let $\mathcal{B}$ a belief base, and $\mathcal{K} = Cn(\mathcal{B})$. The set $\mathcal{K} \operatorname{div} A$ is the base-generated contraction of $\mathcal{K}$ by A iff it satisfies the following postulates:*

BGC1 Closure*: $\mathcal{K} \operatorname{div} A = Cn(\mathcal{K} \operatorname{div} A)$.*
BGC2 Success*: if A is not tautological, then $A \notin Cn(\mathcal{K} \operatorname{div} A)$.*
BGC3 Inclusion*: $\mathcal{K} \operatorname{div} A \subseteq \mathcal{K}$.*
BGC4 Vacuity*: if $A \notin \mathcal{K}$, then $\mathcal{K} \subseteq \mathcal{K} \operatorname{div} A$.*
BGC5 Extensionality*: if A and B are classically equivalent, then $\mathcal{K} \operatorname{div} A = \mathcal{K} \operatorname{div} B$.*
BGC6 Finitude*: there is a finite set $\mathcal{B}$ such that $\mathcal{K} \operatorname{div} A = Cn(\mathcal{B}')$ for some $\mathcal{B}' \subseteq \mathcal{B}$.*
BGC7 Symmetry*: if for all A we have that $B \in \mathcal{K} \operatorname{div} A$ exactly when $C \in \mathcal{K} \operatorname{div} A$, then $\mathcal{K} \operatorname{div} B = \mathcal{K} \operatorname{div} C$.*
BGC8 Conservativity*: if $\mathcal{K} \operatorname{div} B$ is not a subset of $\mathcal{K} \operatorname{div} A$, then there is some C such that $\mathcal{K} \operatorname{div} A \subseteq \mathcal{K} \operatorname{div} C$, $C \notin \mathcal{K} \operatorname{div} C$ and $A \in (\mathcal{K} \operatorname{div} B) \cup (\mathcal{K} \operatorname{div} C)$.*

Proof. For a proof see [11, pp. 610-614]. □

Axioms BGC1 – BGC2 ensure that the result of base-generated contraction of $\mathcal{K}$ by A is a belief set which does not contain A, unless A is tautological: Axioms BGC3 – BGC4 guarantee that such a belief set is identical to $\mathcal{K}$ whenever A does not belong to $\mathcal{K}$. Axiom BGC5 states that $\mathcal{K} \operatorname{div} A$ is not altered by the free substitution of A with a logically equivalent formula, whereas Axiom BGC6 establishes that base-generated contraction of $\mathcal{K}$ by A yields only a finite number of base-generated belief sets – even if A varies over an infinite set of formulas. On the other hand, Axiom BGC7 requires that if there is no contraction by which a sentence B is retracted without another sentence C being retracted, and vice versa, then contraction by B is identical to contraction by C. Lastly, Axiom BGC8 ensures that every element of $\mathcal{K}$ is retained in $\mathcal{K}$ after contraction by A, unless there is some good reason to exclude it[3].

[3] Axiom BGC8 plays for Theorem 4.9 the same role as Axiom BC3 in Definition 4.2 (cf. [11, pp. 605-606] for a detailed discussion).

On this basis, we can give an axiomatic characterization of base-generated revision of a belief set:

Theorem 4.10. *Let $\mathcal{K}$ be a belief set. The set $\mathcal{K} * A$ is the base-generated revision of $\mathcal{K}$ by A iff it satisfies the following postulates:*

BGR1 Closure*: $\mathcal{K} * A = Cn(\mathcal{K} * A)$.*
BGR2 Success*: $A \in \mathcal{K} * A$.*
BGR3 Inclusion*: $\mathcal{K} * A \subseteq \mathcal{K} + A$.*
BGR4 Vacuity*: if $\neg A \notin \mathcal{K}$ then $\mathcal{K} + A \subseteq \mathcal{K} * A$.*
BGR5 Consistency Preservation*: if A is consistent, then $\mathcal{K} * A$ is consistent.*
BGR6 Extensionality*: if A and B are classically equivalent, then $\mathcal{K} * A = \mathcal{K} * B$.*

Proof. By Theorem 4.6, $\mathcal{B} * A = (\mathcal{B} \operatorname{div} \neg A) + A$: this implies that $Cn(\mathcal{B} * A) = Cn((\mathcal{B} \operatorname{div} \neg A) \cup \{A\})$. By reflexivity, monotony and idempotence of classical consequence we have that $Cn((\mathcal{B} \operatorname{div} \neg A) \cup \{A\}) = Cn(Cn(\mathcal{B} \operatorname{div} \neg A) \cup Cn(\{A\})) = Cn((\mathcal{K} \operatorname{div} \neg A) \cup Cn(\{A\})) = Cn((\mathcal{K} \operatorname{div} \neg A) \cup \{A\})$. By the classical result in [1, p. 513], $\mathcal{K} * A$ satisfies Axioms BGR1 – BGR6 if and only if $\mathcal{K} \operatorname{div} \neg A$ satisfies Axioms BGC1 – BGC5: Theorem 4.9 suffices to the conclusion. □

We conclude this section with a classical result concerning the relationship between base-generated belief revision and a weak version of the preferential logic R:

Theorem 4.11. *Let $\mathcal{K}$ be a belief set. Then $\mathcal{K} * A$ satisfies Axioms* BGR1 – BGR6 *and $B \in \mathcal{K} * A$ iff $A \mid B$ for a nonmonotonic consequence relation $\mid$ satisfying the following postulates:*

R1 Right Weakening*: $A \mathrel{|\!\sim} B$ iff $A \mathrel{|\!\sim} C$ and $B \in Cn(\{C\})$.*
R2 Reflexivity*: $A \mathrel{|\!\sim} A$.*
R3 Weak Conditionalization*: If $A \mathrel{|\!\sim} B$, then $\top \mathrel{|\!\sim} A \to B$.*
R4 Weak Rational Monotony*: If $\top \mathrel{|\!\sim} A \to B$ and $\top \not\mathrel{|\!\sim} \neg A$, then $A \mid B$.*
R5 Consistency Preservation*: If A is consistent, then there is (at least) one formula B such that $A \not\mathrel{|\!\sim} B$.*
R6 Left Logical Equivalence*: If A and B are classically equivalent, then $A \not\mathrel{|\!\sim} B$ iff $A \mathrel{|\!\sim} C$.*

Proof. For a proof see [17]. □

5. *Hypersequent calculi for base-generated belief revision*

5.1. *Hybrid hypersequents*

Hypersequents are denoted by capital Latin letters $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2, \ldots$. In proof-theoretic literature, a hypersequent is a list of sequents which are separated by the bar symbol '|': e.g., $\mathcal{H} = \Gamma_1 \vdash \Delta_1 \mid \cdots \mid \Gamma_n \vdash \Delta_n$. Usually, hypersequents are taken to be multisets instead of lists: e.g., $\mathcal{H} = \Gamma_{\pi(1)} \vdash \Delta_{\pi(1)} \mid \cdots \mid \Gamma_{\pi(n)} \vdash \Delta_{\pi(n)}$ for any permutation π on $1, \ldots, n$. Now the intended interpretation of a hypersequent $\Gamma_1 \vdash \Delta_1 \mid \cdots \mid \Gamma_n \vdash \Delta_n$ is that $\Gamma_i \vdash \Delta_i$ is provable, for some $1 \leq i \leq n$. In other words, the bar symbol '|' receives an interpretation in terms of disjunction: if the formula translation of a sequent $\Gamma \vdash \Delta$ is $\bigwedge \Gamma \rightarrow \bigvee \Delta$, then the formula translation of the hypersequent $\Gamma_1 \vdash \Delta_1 \mid \cdots \mid \Gamma_n \vdash \Delta_n$ is $\bigvee\limits_{i=1}^{n} (\bigwedge \Gamma_i \rightarrow \bigvee \Delta_i)$.

In this paper, we employ a non-standard notion of hypersequent, whereby a hypersequent is a *set* made up of a sequent $\Gamma \vdash \Delta$ and a (possibly, empty) set of antisequents $\Gamma_1 \dashv \Delta_1, \cdots, \Gamma_n \dashv \Delta_n$: e.g., $\mathcal{H} = \Gamma_1 \vdash \Delta_1 \mid \Gamma_2 \dashv \Delta_2 \mid \cdots \mid \Gamma_n \dashv \Delta_n$ is a hypersequent, with either $\Gamma_i \neq \Gamma_j$ or $\Delta_i \neq \Delta_j$ for any $1 \leq j \neq i \leq n$, and with $\mathcal{H} = \Gamma_1 \vdash \Delta_1 \mid \Gamma_{\pi(2)} \dashv \Delta_{\pi(2)} \mid \cdots \mid \Gamma_{\pi(n)} \dashv \Delta_{\pi(n)}$ for any permutation π defined on $2, \ldots, n$. The intended interpretation of a hypersequent $\Gamma_1 \vdash \Delta_1 \mid \Gamma_2 \dashv \Delta_2 \mid \cdots \mid \Gamma_n \dashv \Delta_n$ is that $\Gamma \vdash \Delta$ is provable and $\Gamma_k \vdash \Delta_k$ is refutable, for any $2 \leq k \leq n$: in this case, the bar symbol '|' is *metalogically* interpreted as conjunction. With both sequents and antisequents present, our hypersequents do not have formula translations: they serve as syntactical objects expressing *contrary updating* on the provability of the included sequents. For any hypersequent $\mathcal{H}$, we denote the sets of antisequents in $\mathcal{H}$ with capital Latin letters $\mathcal{R}, \mathcal{R}_1, \mathcal{R}_2, \ldots \mathcal{S}, \mathcal{S}_1, \mathcal{S}_2$. Specifically, in any hypersequent $\mathcal{H}$ of the form $\Gamma \vdash \Delta \mid \mathcal{R}$, the letter $\mathcal{R}$ designates the *refutational part* of $\mathcal{H}$.

5.2. *A proof-theoretic approach to base revision*

Let $\mathcal{B}$ be a belief base, and B be the conjunction of the formulas in $\mathcal{B}$. We exploit Proposition 2.3 to define a unique, logically equivalent set $\mathcal{B}'$ comprising the formula translations of the clauses in $\mathsf{top_c}(\overset{*}{\vdash} B)$: in what follows, we shall always consider $\mathcal{B}$ to be identical to such $\mathcal{B}'$. This assumption allows for a fine-grained control over the information stored in agent's beliefs, while making the notion of belief base less intensional: we deem that the price of this move is worth paying, at least if we refrain from adhering to a foundationalist view of the agent's doxastic apparatus [12].

In our framework, the result of the base expansion of $\mathcal{B}$ by a consistent formula A is represented by the set of clauses $\mathsf{top}(\overset{*}{\vdash} B \wedge A)$. Theorem 3.7 allows us to pinpoint the mcs's $\mathcal{C}_1, \ldots, \mathcal{C}_m$ of $\mathsf{top}(\vdash B \wedge A)$ which include $\mathsf{top_c}(\vdash A)$: for any $1 \leq i \leq m$, $\mathcal{C}_i$ is a subset of $\mathsf{top}(\vdash B \wedge A)$ including $\mathsf{top_c}(\vdash A)$ and such that $\{\vdash p_i\}_{1 \leq i \leq m}$ and $\{q_j \vdash\}_{1 \leq j \leq n}$ belong to $\mathcal{C}_i^\star$ for any clause $p_1, \ldots, p_m \vdash q_1, \ldots, q_n$ in $\mathsf{top}(\vdash B \wedge A) \setminus \mathcal{C}_i$. On the other hand, given a selection function γ for $\mathcal{B}$, Proposition 4.7 establishes that the result of the base revision of $\mathcal{B}$ by A is $\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_i$, for some

$1 \leq i \leq m$: this yields a syntactic characterization of the result of the base revision of $\mathcal{B}$ by A.

Example 5.1. Let $\mathcal{B}$ be the set $\{p, q \to r, \neg q, \neg r\}$ and A be $p \to q$. By Theorem 3.7, the mcs's of $\mathsf{top}(\vdash B \wedge A)$ which include $\mathsf{top_c}(\vdash A)$ are the following:

$$\begin{aligned} (i)\ & \mathcal{C}_1 = \{p \vdash q;\ \vdash p;\ q \vdash r\}; \\ (ii)\ & \mathcal{C}_2 = \{p \vdash q;\ \vdash p;\ r \vdash\}; \\ (iii)\ & \mathcal{C}_3 = \{p \vdash q;\ q \vdash;\ r \vdash;\ q \vdash r\}. \end{aligned}$$

If $\gamma(\mathcal{B} \perp A) = \{\mathcal{C}_1, \mathcal{C}_2\}$, then the result of the base revision of $\mathcal{B}$ by A based on γ is $\mathcal{C}_1 \cap \mathcal{C}_2 = \{p \vdash q;\ \vdash p\}$.

Now, consider $\mathcal{K} {*} A = Cn(\mathcal{B} {*} A)$: by Propositions 2.8 and 3.1, $B \in \mathcal{K} *A$ exactly when the $\overline{\overline{\mathsf{G4str}}}$ calculus for $\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_i$ proves $\vdash B$. By Lemma 3.6 and Proposition 3.5, a clause $\Theta \vdash \Lambda$ can be derived from $\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_i$ if and only if there is a $\overline{\overline{\mathsf{G4str}}}$-proof of $\Theta \vdash \Lambda$ where a (possibly, empty) sequence of Cut applications is followed by a (possibly, empty) sequence of Weakening applications: as a result, $\Theta \vdash \Lambda$ is a (possibly, weakened) clause $\Phi \vdash \Psi$ in $(\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_i)^\star$. On the other hand, $(\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_i)^\star$ is identical to the set $\mathcal{C}_1^\star \cap \cdots \cap \mathcal{C}_i^\star$ (cf. the proof of Proposition 3.11): this implies that the set of clauses which are classically derivable from $\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_i$ is just the closure under Weakening of the set $\mathcal{C}_1^\star \cap \cdots \cap \mathcal{C}_i^\star$. In the next subsection we leverage this fact to obtain hypersequent calculi which are adequate with respect to the base-generated revision of $\mathcal{K}$ by A.

Example 5.2. Let $\mathcal{B}$, A and γ be as in Example 5.1. We have that:

$$\begin{aligned} (i)\ & \mathcal{C}_1^\star = \{p \vdash q;\ \vdash p;\ q \vdash r;\ \vdash q;\ \vdash r;\ p \vdash r\}; \\ (ii)\ & \mathcal{C}_2^\star = \{p \vdash q;\ \vdash p;\ r \vdash;\ \vdash q\} \end{aligned}$$

The closure under Cut of the result of the base revision of $\mathcal{B}$ by A based on γ is $(\mathcal{C}_1 \cap \mathcal{C}_2)^\star = \mathcal{C}_1^\star \cap \mathcal{C}_2^\star = \{p \vdash q;\ \vdash p;\ \vdash q\}$. It's worth noting that the mcs's of $\mathsf{top}^\star(\vdash B \wedge A)$ including $\mathsf{top_c}(\vdash A)$ are $\mathcal{C}_1^\star$, $\mathcal{C}_2^\star$, $\mathcal{C}_3^\star$ and the set $\{p \vdash q;\ q \vdash;\ \vdash r;\ q \vdash r;\ p \vdash r;\ p \vdash\}$.

5.3. *Hypersequent calculi for belief revision*

Let $\mathcal{B}$ be a belief base, with B being the conjunction of its formulas, and A be a consistent formula. Moreover, let $\mathcal{C}_1, \ldots, \mathcal{C}_n$ be the mcs's of $\mathsf{top}(\vdash B \wedge A)$ which include $\mathsf{top}(\vdash A)$. Given a selection function γ for $\mathcal{B}$, the hypersequent calculus $\mathsf{HG4}$ for $\mathcal{K} * A$ based on γ has the rules displayed in Figure 2, with the *proviso* that each instance of ax fulfills the following side conditions:

(i) either $\Theta \cap \Lambda \neq \varnothing$ or $\Theta \overset{*}{\models} \Lambda$ is a clause in $\mathsf{top}_\mathsf{c}^\star(\overset{*}{\models} B \wedge A)$;

(ii) the refutational part $\mathcal{R}$ is $\Theta_1 \dashv \Lambda_1 \mid \cdots \mid \Theta_k \dashv \Lambda_k$, with $\{\Theta_h \mathrel{|\!\sim}^{*} \Lambda_h\}_{1 \leq h \leq k} = \mathsf{top}^{\star}_{\mathsf{c}}(\mathrel{|\!\sim}^{*} B \wedge A) \setminus \{\mathcal{C}^{\star}_1 \cap \cdots \cap \mathcal{C}^{\star}_i\}$ for some $1 \leq i \leq n$.

AXIOMS

$$\frac{}{\Gamma, \Theta \vdash \Lambda, \Delta \mid \mathcal{R}}\, ax$$

LOGICAL RULES

$$\frac{\Gamma \vdash \Delta, A \mid \mathcal{R}}{\Gamma, \neg A \vdash \Delta \mid \mathcal{R}}\, \neg_L \qquad \frac{\Gamma, A \vdash \Delta \mid \mathcal{R}}{\Gamma \vdash \Delta, \neg A \mid \mathcal{R}}\, \neg_R$$

$$\frac{\Gamma, A, B \vdash \Delta \mid \mathcal{R}}{\Gamma, A \wedge B \vdash \Delta \mid \mathcal{R}}\, \wedge_L \qquad \frac{\Gamma \vdash \Delta, A \mid \mathcal{R}_1 \qquad \Gamma \vdash \Delta, B \mid \mathcal{R}_2}{\Gamma \vdash \Delta, A \wedge B \mid \mathcal{R}_1 \mid \mathcal{R}_2}\, \wedge_R$$

$$\frac{\Gamma, A \vdash \Delta \mid \mathcal{R}_1 \qquad \Gamma, B \vdash \Delta \mid \mathcal{R}_2}{\Gamma, A \vee B \vdash \Delta \mid \mathcal{R}_1 \mid \mathcal{R}_2}\, \vee_L \qquad \frac{\Gamma \vdash \Delta, A, B \mid \mathcal{R}}{\Gamma \vdash \Delta, A \vee B \mid \mathcal{R}}\, \vee_R$$

$$\frac{\Gamma \vdash \Delta, A \mid \mathcal{R}_1 \qquad \Gamma, B \vdash \Delta \mid \mathcal{R}_2}{\Gamma, A \to B \vdash \Delta \mid \mathcal{R}_1 \mid \mathcal{R}_2}\, \to_L \qquad \frac{\Gamma, A \vdash \Delta, B \mid \mathcal{R}}{\Gamma \vdash \Delta, A \to B \mid \mathcal{R}}\, \to_R$$

FIGURE 2 **HG4** hypersequent calculi for $\mathcal{K} * A$.

Example 5.3. Let $\mathcal{B}$, A and γ be as in Example 5.1. The **HG4** calculus for $\mathcal{K} * A$ based on γ features

$$\Gamma, p \vdash q, \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r | p \dashv r | \dashv \Gamma, q \vdash r, \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r | p \dashv r | \dashv$$

$$\Gamma, p \vdash r, \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r \mid p \dashv r | \dashv \Gamma \vdash \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r | p \dashv r | \dashv$$

$$\Gamma \vdash p, \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r | p \dashv r | \dashv \Gamma \vdash q, \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r | p \dashv r | \dashv$$

$$\Gamma, p \vdash \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r | p \dashv r | \dashv \Gamma, q \vdash \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r | p \dashv r | \dashv$$

$$\Gamma \vdash r, \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r | p \dashv r | \dashv \Gamma, r \vdash \Delta | p \dashv | q \dashv | \dashv r | r \dashv | q \dashv r | p \dashv r | \dashv$$

as initial hypersequents.

It is easy to find cases where there is (at least) one derivation of a hypersequent $\Gamma \vdash \Delta \mid \mathcal{R}$ in the **HG4** calculus for $\mathcal{K} * A$ based on a selection function γ such that $\bigwedge \Gamma \to \bigvee \Delta$ does not belong to $\mathcal{K} * A$. To address this problem, we distinguish **HG4**-provability from **HG4**-derivability along the following lines:

Definition 5.4. A **HG4**-derivation π is a *proof* if $\overline{\overline{\mathsf{G4}}}$ refutes $\bigwedge \Theta \to \bigvee \Lambda \vdash \bigwedge \Theta_i \to \bigvee \Lambda_i$ for any instance of ax in π with conclusion $\Gamma, \Theta \vdash \Lambda, \Delta \mid \Theta_1 \dashv \Lambda_1 \mid \cdots \mid \Theta_n \dashv \Lambda_n$ and each $1 \leq i \leq n$, and a *paraproof* otherwise.

Example 5.5. Let $\mathcal{B}$, A and γ be as in Example 5.1. The following HG4-derivation is a proof:

$$\dfrac{\neg_L\,\dfrac{ax\,\dfrac{}{\vdash p, r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv}}{\neg p \vdash r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv} \qquad \dfrac{\dfrac{}{\vdash q, r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv}\,ax}{\neg q \vdash r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv}\,\neg_L}{\neg p \vee \neg q \vdash r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv}\,\vee_L$$

On the other hand, the following HG4-derivation is a paraproof:

$$\dfrac{\neg_L\,\dfrac{ax\,\dfrac{}{\vdash p, r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv}}{\neg p \vdash r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv} \qquad \dfrac{\dfrac{}{\vdash r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv}\,ax}{\neg r \vdash r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv}\,\neg_L}{\neg p \vee \neg r \vdash r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv}\,\vee_L$$

In the remaining part of this section we exploit the structural properties of HG4 calculi to show that HG4-provability is sound and (weakly) complete with respect to base-generated belief revision. Let us begin by stating that a rule of the form

$$\dfrac{\Gamma_1 \vdash \Delta_1 \mid \mathcal{R}_1 \qquad \cdots \qquad \Gamma_n \vdash \Delta_n \mid \mathcal{R}_n}{\Gamma \vdash \Delta \mid \mathcal{R}_1 \mid \cdots \mid \mathcal{R}_n}\,r$$

is *admissible* in HG4 if the hypersequent $\Gamma \vdash \Delta \mid \mathcal{R}_1 \mid \cdots \mid \mathcal{R}_n$ is provable whenever the hypersequents $\{\Gamma_i \vdash \Delta_i \mid \mathcal{R}_i\}_{1 \leq i \leq n}$ are provable, and *absorbed* if the hypersequent $\Gamma \vdash \Delta \mid \mathcal{R}_1 \mid \cdots \mid \mathcal{R}_n$ is derivable whenever the hypersequents $\{\Gamma_i \vdash \Delta_i \mid \mathcal{R}_i\}_{1 \leq i \leq n}$ are derivable.

Proposition 5.6. *The rules of Left and Right Weakening*

$$wk\,\dfrac{\Gamma \vdash \Delta \mid \mathcal{R}}{A, \Gamma \vdash \Delta \mid \mathcal{R}} \qquad\qquad \dfrac{\Gamma \vdash \Delta \mid \mathcal{R}}{\Gamma \vdash \Delta, A \mid \mathcal{R}}\,wk$$

are height-preserving admissible in HG4.

Proof. We reason by routine induction on the height of a HG4-proof π of $\Gamma \vdash \Delta \mid \mathcal{R}$: as usual, the height of π is taken to be the number of nodes in a branch of maximal length. □

Corollary 5.7. *The rules of Left and Right Weakening can be absorbed in* HG4 *with preservation of height.*

Theorem 5.8. *The rule of Cut*

$$\frac{\Gamma \vdash \Delta, A \mid \mathcal{R}' \qquad A, \Pi \vdash \Sigma \mid \mathcal{R}''}{\Pi, \Gamma \vdash \Delta, \Sigma \mid \mathcal{R}' \mid \mathcal{R}''}\, cut$$

is admissible in HG4.

Proof. We argue as for the proof of Theorem 2.7. □

Corollary 5.9. *The rule of Cut can be absorbed in* HG4.

Proposition 5.10. *If there exists a* HG4*-proof π of $\Gamma \vdash \Delta \mid \mathcal{R}$, then for any hypersequent $\Gamma' \vdash \Delta' \mid \mathcal{R}'$ in π the formulas in Γ', Δ' are subformulas of formulas in Γ, Δ.*

Proof. By induction on the height of π. □

We say that a rule of the form

$$\frac{\Gamma_1 \vdash \Delta_1 \mid \mathcal{R}_1 \qquad \cdots \qquad \Gamma_n \vdash \Delta_n \mid \mathcal{R}_n}{\Gamma \vdash \Delta \mid \mathcal{R}_1 \mid \cdots \mid \mathcal{R}_n}$$

is *invertible* if and only if a rule of the form

$$\frac{\Gamma \vdash \Delta \mid \mathcal{R}_1 \mid \cdots \mid \mathcal{R}_n}{\Gamma_i \vdash \Delta_i \mid \mathcal{R}_1 \mid \cdots \mid \mathcal{R}_n}$$

is admissible in HG4, for any $1 \leq i \leq n$.

Proposition 5.11. *Logical rules of* HG4 *are height-preserving invertible.*

Proof. By routine induction on the height of the HG4-proofs of hypersequents $B \circ C, \Gamma \vdash \Delta \mid \mathcal{R}$ and $\Gamma \vdash \Delta, B \circ C \mid \mathcal{R}$, with $\circ \in \{\wedge, \vee, \rightarrow\}$, as well as hypersequents $\neg B, \Gamma \vdash \Delta \mid \mathcal{R}$ and $\Gamma \vdash \Delta, \neg B \mid \mathcal{R}$. □

Theorem 5.12. *A hypersequent $\Gamma \vdash \Delta \mid \mathcal{R}$ is provable in the* HG4 *calculus for $\mathcal{K} * A$ based on a selection function γ iff $\bigwedge \Gamma \rightarrow \bigvee \Delta$ belongs to $\mathcal{K} * A$ based on γ.*

Proof. ($\Rightarrow$) We reason by induction on the height of a HG4-proof π of $\Gamma \vdash \Delta \mid \mathcal{R}$. If $h(\pi) = 1$, then π is of the form

$$\frac{}{\Gamma, \Theta \vdash \Lambda, \Delta \mid \Theta_1 \dashv \Lambda_1 \mid \cdots \mid \Theta_k \dashv \Lambda_k}\, ax$$

Definition 5.4 ensures that $\overline{\overline{\mathsf{G4}}}$ refutes $\bigwedge \Theta \rightarrow \bigvee \Lambda \vdash \bigwedge \Theta_i \rightarrow \bigvee \Lambda_i$, for any $1 \leq i \leq n$: as a result, $\Theta \vdash \Lambda$ does not belong to $\mathsf{top}^{\star}_{\mathsf{c}}(\vdash B \wedge A)\setminus$

$(\mathcal{C}_1^\star \cap \cdots \cap \mathcal{C}_i^\star)$ – on pain of contradiction. This implies that $\bigwedge \Theta \to \bigvee \Lambda$ and thus $(\bigwedge \Gamma \wedge \bigwedge \Theta) \to (\bigvee \Lambda \vee \bigvee \Delta)$ belong to $\mathcal{K} * A$. The inductive step is obvious – and we are done.

($\Leftarrow$) We reason by induction on the length of a Hilbert-style deduction δ of $\bigwedge \Gamma \to \bigvee \Delta$ from a set of axioms for classical propositional logic along with the set of extra-logical axioms in (refined) $\mathcal{B} * A$.

[BASE] If $lh(\delta) = 1$ and $\bigwedge \Theta \to \bigvee \Lambda$ is an extra-logical axiom, then suppose by contradiction that $\overline{\overline{\mathsf{G4}}}$ proves $\bigwedge \Theta \to \bigvee \Lambda \vdash \bigwedge \Theta_i \to \bigvee \Lambda_i$ for some $1 \leq i \leq n$. By Proposition 3.5, this implies that the clause $\Theta_i \vdash \Lambda_i$ is a weakened version of $\Theta \vdash \Lambda$. On the other hand, $\Theta \vdash \Lambda$ belongs to some mcs $\mathcal{C}$ of $\mathsf{top}(\vdash B \wedge A)$ which does not contain $\Theta_i \vdash \Lambda_i$: if $\Theta_i = p_1, \ldots, p_m$ and $\Lambda_i = q_1, \ldots, q_n$, Theorem 3.7 ensures that the clauses $\{\vdash p_h\}_{1 \leq h \leq m}$ and $\{q_k \vdash\}_{1 \leq k \leq n}$ belong to $\mathcal{C}^\star$. Since $\Theta \subseteq \Theta_i$ and $\Lambda \subseteq \Lambda_i$, by Lemma 3.6 we would get a contradiction: we can thus infer that $\overline{\overline{\mathsf{G4}}}$ refutes $\bigwedge \Theta \to \bigvee \Lambda \vdash \bigwedge \Theta_i \to \bigvee \Lambda_i$ for any $1 \leq i \leq n$, as desired.

[STEP] If $lh(\delta) \geq n + 1$, then the last rule applied is modus ponens: we exploit Proposition 5.11 and Theorem 5.8 to reach the conclusion. □

Corollary 5.13. *A hypersequent $\Gamma \vdash \Delta \mid \mathcal{R}$ is provable in the* HG4 *calculus for $\mathcal{K} * A$ based on a selection function γ iff $A \Big| (\bigwedge \Gamma \to \bigvee \Delta)$ for a nonmonotonic consequence relation $\Big|$ satisfying Axioms* R1 – R6.

Proof. Straightforward from Theorems 5.12 and 4.11. □

Theorem 5.12 establishes that $\vdash A \mid \mathcal{R}$ is provable in HG4 if A belongs to the base-generated revision of $\mathcal{K}$ by A, for some selection function γ for $\mathcal{B}$. However, one cannot prove the stronger claim that $\Gamma \vdash A \mid \mathcal{R}$ is provable in HG4 if A belongs to the base-generated revision of $\mathcal{K} \cup \Gamma$ by A, for the same selection function γ. In short, HG4 calculi fail to be strongly complete with respect to base-generated belief revision.

Example 5.14. Let $\mathcal{B}$, A and γ be as in Example 5.1. As shown in Example 5.5, the hypersequent $\neg p \vee \neg q \vdash r \mid p \dashv \mid q \dashv \mid \dashv r \mid r \dashv \mid q \dashv r \mid p \dashv r \mid \dashv$ is provable: by Theorem 5.12, this means that $(\neg p \vee \neg q) \to r$ belongs to the revision of $\mathcal{K}$ by A based on γ. On the other hand, Theorem 3.7 entails that the mcs's of $\mathsf{top}(\vdash B \wedge (\neg p \vee \neg q) \wedge A)$ which include $\mathsf{top_c}(\vdash A)$ are the following:

(i) $\mathcal{D}_1 = \{p \vdash q;\ p, q \vdash;\ q \vdash r;\ q \vdash;\ r \vdash\}$
(ii) $\mathcal{D}_2 = \{p \vdash q;\ \vdash p;\ q \vdash r\}$
(iii) $\mathcal{D}_3 = \{p \vdash q;\ \vdash p;\ r \vdash\}$.

If either $\gamma((\mathcal{B} \cup \{\neg p \vee \neg q\})\bot A) = \mathcal{D}_i \cap \mathcal{D}_j$ for any $1 \leq i \neq j \leq 3$, or $\gamma((\mathcal{B} \cup \{\neg p \vee \neg q\})\bot A) = \mathcal{D}_i$ for $i = 1, 3$, or $\gamma((\mathcal{B} \cup \{\neg p \vee \neg q\})\bot A) = \mathcal{D}_1 \cap \mathcal{D}_2 \cap \mathcal{D}_3$, then r does not belong the revision of $\mathcal{K} \cup \{\neg p \vee \neg q\}$ by A.

6. *Conclusion*

In this paper, we leverage a syntactic account of maximal consistent subsets of sets of clauses to define hybrid hypersequent calculi for a refined version of AGM belief revision. Specifically, we exploit the parallel composition of sequents and antisequents built into hybrid hypersequents to express contrary updating on the provability of initial sequents: this feature allows us to give an adequate formalization of base-generated revision – equivalently, a weak version of system R.

For future work, we believe that a hypersequent-based approach to iterated AGM belief revision would be promising, especially in view of the largely open problem of finding adequate operators satisfying DP postulates [5]. Moreover, we deem that hypersequent calculi for AGM belief revision can be easily adapted to multiple bases revision, with possible applications to belief merging [8].

This work represents a step forward towards the realization of a broader goal: providing a uniform proof-theoretic platform for monotonic and nonmonotonic extensions of classical propositional logic by means of combinations of sequents and antisequents framed in suitable Gentzen-style calculi [19, 22, 24].

In this perspective, it would be valuable to extend the scope of application of the hypersequent-based approach to include KLM logics [15]. If we confine ourselves to KLM logics formulated in finite propositional languages, it appears that sound and (weakly) complete HG4 calculi can be obtained by endowing hybrid hypersequents with a tree-like structure: the binary relation $<$ on states would be codified by the refutational part of initial hypersequents, whereas formal conditions on $<$ would be captured by suitable structural rules.

Finally, it would be interesting to compare the expressive power of HG4 calculi with the scope of application displayed by adaptive logics [28] and controlled calculi [20].

References

[1] C. E. Alchourrón, P. Gärdenfors and D. Makinson, *On the logic of theory change: partial meet contraction and revision functions*, J. Symb. Log. **50** (1985), 510–530.

[2] A. Avron, *Hypersequents, logical consequence and intermediate logics for concurrency*, Ann. Math. Artif. Intell. **4** (1991), 225–248.

[3] A. Avron, *Gentzen-type systems, resolution and tableaux*, J. Automat. Reason. **10** (1993), 265–281.

[4] A. Avron, *The method of hypersequents in the proof theory of propositional non-classical logics*, In: "Logic: Foundations to Applications", W. Hodges (ed.), Oxford University Press, 1996.

[5] A. Darwiche and J. Pearl, *On the logic of iterated belief revision*, Artificial Intelligence **89** (1997), 1–29.

[6] V. Goranko, *Refutation systems in modal logic*, Studia Logica, **53** (1994), 299–324.

[7] V. Goranko, G. Pulcini and T. Skura, "Refutation systems: an overview and some applications to philosophical logics", pages 173–197. Springer Berlin Heidelberg, 2020.

[8] E. Grégoire and S. Konieczny, *Logic-based approaches to information fusion*, Inf. Fusion Data Sci. **7** (2006), 4–18.

[9] S. O. Hansson, *Knowledge-level analysis of belief base operations*, Artificial Intelligence **82** (1991), 215–235.

[10] S. O. Hansson, *Reversing the Levi identity*, J. Philos. Logic **22** (1993), 637–669.

[11] S. O. Hansson, *Theory contraction and base contraction unified*, J. Symb. Log. **58** (1993), 602–625.

[12] G. Harman, "Change in View", MIT Press, 1986.

[13] B. Hill and F. Poggiolesi, *An analytic calculus for the intuitionistic logic of proofs*, Notre Dame J. Form. Log. **60** (2019), 353–393.

[14] S. Kleene, *Mathematical Logic*, John Wiley and Sons, 1967.

[15] S. Kraus, D. Lehmann and M. Magidor, *Nonmonotonic reasoning, preferential models and cumulative logics*, Artificial Intelligence **44** (1990), 167–207.

[16] I. Levi, "The Fixation of Belief and its Undoing", Cambridge University Press, 1991.

[17] D. Makinson and P. Gärdenfors, *Relations between the logic of theory change and nonmonotonic logic*, In: "The Logic of Theory Change", A. Fuhrmann and M. Morreau (eds.), Oxford University Press, 1991.

[18] S. Negri and J. von Plato, *Cut elimination in sequent calculi with implicit contraction, with a conjecture on the origin of Gentzen's altitude construction*, In: "Concepts of Proof in Mathematics, Philosophy and Computer Science", D. Probst and P. Schuster (eds.), de Gruyter, 2016.

[19] M. Piazza and G. Pulcini, *Uniqueness of axiomatic extensions of cut-free classical propositional logic*, Log. J. IGPL **24** (2016), 708–718.

[20] M. Piazza and G. Pulcini, *Unifying logics via context-sensitiveness*, J. Logic Comput. **27** (2017), 21–40.

[21] M. Piazza and G. Pulcini, *Fractional semantics for classical logic*, Rev. Symb. Log. **13** (2020), 810–828.

[22] M. Piazza, G. Pulcini and A. Sabatini, *Abduction as deductive saturation: a proof-theoretic inquiry*, J. Philos. Logic **52** (2023), 1575–1602.

[23] M. Piazza, G. Pulcini and M. Tesi, *Linear logic in a refutational setting*, J. Logic Comput. **52** (2023), 1–25.

[24] M. Piazza and A. Sabatini, *Hypersequent calculi for propositional default logics*, ACM Transactions on Computational Logic, forthcoming, https://doi.org/10.1145/3725849.

[25] M. Piazza and A. Sabatini, *On anticut rules: classical, FDE-based and intuitionistic logics*, The Australasian Journal of Logic, forthcoming.

[26] M. Piazza and M. Tesi, *Analyticity with extra-logical information*, J. Logic Comput. (2024).

[27] G. Pulcini and A. C. Varzi, *Classical logic through refutation and rejection*, In: "Landscapes in Logic (Volume on Philosophical Logics)", M. Fitting (ed.), College Publications, 2021.

[28] C. Straßer, "Adaptive Logics for Defeasible Reasoning. Applications in Argumentation, Normative Reasoning and Default Reasoning", Trends in Logic. Springer, 2013.

Andrea Sabatini

Belief Structures within Fractional Semantics: an overview

1. *Introduction*

Fractional Semantics, initially introduced in [17], serves as a powerful tool for discerning the number of proper axioms within a proposition relative to the total number of axioms. This method underwent refinement for modal logic [19] and expanded into the domain of beliefs in [3] and applied to the Lottery Paradox in [2]. The study demonstrated the instrumental role of Fractional Semantics in resolving the Lottery Paradox.

This work has two main objectives: firstly, to present in a refined way $GS4_B$, firstly presented in [3]—the Fractional Semantics System that incorporates beliefs; secondly, to introduce a nuanced categorization of beliefs. In [3], all beliefs are treated as true, akin to tautologies. However, this poses a philosophical challenge, as not every proposition we believe aligns with the certainty of a tautology. To address this, we utilize Hyperreal numbers, signifying that a belief holds a value not of 1, but infinitesimally lower—specifically, $1 - \delta$, where δ represents an infinitesimal value smaller than every real number.

This approach draws inspiration from Hansson [9,10], who used hyperreal numbers to differentiate between Full Beliefs (assigned a value of 1) and beliefs open to revision in the presence of evidence, termed Revisable Beliefs. However, our aim is different: we seek a system capable of tracking not only the count of Full Beliefs but also beliefs considered true even if subject to revision, differenciating between them thanks to hyperreal numbers. Fractional Semantics enables us to perform derivations and determine the composition of the combination between tautologies, Full Beliefs, and Revisable Beliefs.

The paper is structured as follows: in the first section, we briefly present Fractional Semantics, referring to [2, 3, 17–19] for more examples and proofs; in the second section, we present proofs for theorems from [3]; and in the last section, we introduce a distinction between Full Beliefs and Revisable Beliefs within the framework of fractional semantics.

2. *Fractional Semantics*

Fractional semantics is a multi-valued approach governed by pure proof-theoretic considerations firstly introduced in [17], assigning truth values as rational numbers in the closed interval [0,1] breaking the symmetry between tautologies and contradictions, allowing values other than 0 for non-logical axioms, *i.e.*, contingent. It measures the proposition's proximity to being a tautology or a contradiction.

To enable fractional interpretation, a decidable logic $\mathcal{L}$ is required, displayed in a sequent system **S** meeting three conditions: bilateralism, invertibility, and stability.

Bilateralism: **S**, as a bilateral system, generates **S**-derivations for any well formed formula *A* of $\mathcal{L}$: if *A* is valid, its **S**-derivation will be an actual proof of *A*; if *A* is invalid, its **S**-derivation will provide a formal refutation of *A*, *i.e.*, a proof of its unprovability.

Invertibility: each logical rule of *S* is invertible, meaning that the provability of its conclusion implies the provability of (each one of) its premise(s). This means that there is an algorithm to decompose uniquely a sequent into an equivalent formula in conjunctive normal form.

Stability: two analytic S-proofs with the same end-sequent share the same multi-set of top-sequents.

Fractional semantics is obtained by focusing on the axiomatic structure of proofs expressed in Kleene's one-side sequent system $GS4$ [13, 22]. The system is as following:

$$\frac{}{\vdash \Gamma, p, \overline{p}}\ (ax.)$$

$$\frac{\vdash \Gamma, p, q}{\vdash \Gamma, p \vee q}\ (\vee) \qquad\qquad \frac{\vdash \Gamma, p \qquad \vdash \Gamma, q}{\vdash \Gamma, p \wedge q}\ (\wedge)$$

GS4 is a one-sided sequent where structural properties are absorbed into the calculus, Γ and Δ are multisets of formulas, and $p, q, \ldots$ are atomic formulas. As usual, $\wedge$ indicates the conjunction and $\vee$ the disjunction. There is not a rule governing negation as it is inductively defined by different atomic formulas p and $\overline{p}$, where $\overline{p}$ indicates the negation of p. Sequents can be decomposed into initial sequents that are allowed to contain only atomic formulas.

The interpretation of a formula is the result of the ratio between the number of identity top-sequents $(\Delta, p, \overline{p})$ out of the total number of top-sequents occurring in any of its proofs. Weakening and contraction are dropped while cut rule has the form:

$$\frac{\vdash \Gamma, p \qquad \vdash \overline{p}, \Delta}{\vdash \Gamma, \Delta}\ (cut)$$

In order to give a fractional interpretation a counterpart is needed, namely $\overline{\overline{GS4}}$, that is the $GS4$ calculus maximally extended:

Definition 2.1 ($\overline{\overline{GS4}}$)**.** The calculus $\overline{\overline{GS4}}$ is obtained from $GS4$ that is able to prove any sequent and it satisfies cut-elimination à la Gentzen if its axioms introduce only clauses [17], *i.e.*, a sequent which consists solely of atomic formulae [1].

Definition 2.2 (Top-sequents axioms).

$top^1(\pi)$: denotes the multiset of all and only π's *top-sequents* introduced by an identity axiom, *i.e.*, those those sequents directly introduced as instances of the axiom rules.;

$top^0(\pi)$: denotes the multiset of all and only π's *top-sequents* introduced by a complementary axiom, in other words, those axioms that are not tautological.

Any formula A can be interpreted as the ratio between the number of identity top-sequents (sequents introduced by the standard axiom) out of the total number of top-sequents.

$$[\![A]\!] = \frac{top^1(\pi)}{top^1(\pi) + top^0(\pi)}$$

Definition 2.3 (Top-sequents). Top-sequents represent the number of the leaves of the proof as defined in Definition 2.2 and $[\![\Gamma]\!]$ denotes the value of the formula $\vee\Gamma$ where only $\vee-$applications appear.

$top^1(\pi)$: let's call this *m*;
$top^0(\pi)$: let's call this *n*;
$[\![\vee\Gamma]\!]$ is $\frac{m}{n} \in [0,1]$.

From this definition it is possible to give general rules with decorated sequents. These decorated sequents are able to keep track of the fractional value along the proof.

$$\frac{}{\vdash^{1}_{1} \Gamma, p, \overline{p}}\ (ax.) \qquad\qquad \frac{}{\vdash^{0}_{1} \Delta}\ (\overline{ax.})$$

$$\frac{\vdash^{m}_{\mathrm{n}} \Gamma, A, B}{\vdash^{m}_{\mathrm{n}} \Gamma, A \vee B}\ (\vee) \qquad\qquad \frac{\vdash^{m_1}_{\mathrm{n}_1} \Gamma, A \qquad \vdash^{m_2}_{\mathrm{n}_2} \Gamma, B}{\vdash^{m_1+m_2}_{\mathrm{n}_1+\mathrm{n}_2} \Gamma, A \wedge B}\ (\wedge)$$

Example 2.4. Let's consider an example with the turnstile decorated:

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\vdash^{0}_{1} p, q}\,(\overline{ax.})}{\vdash^{0}_{1} p \vee q}\,(\vee) \qquad \dfrac{\dfrac{}{\vdash^{1}_{1} p, \overline{p}}\,(ax.)}{\vdash^{1}_{1} p \vee \overline{p}}\,(\vee)}{\vdash^{1}_{2} (p \vee q) \wedge (p \vee \overline{p})}\,(\wedge) \qquad \dfrac{\dfrac{}{\vdash^{0}_{1} \overline{r}}\,(\overline{ax.}) \qquad \dfrac{}{\vdash^{0}_{1} \overline{t}}\,(\overline{ax.})}{\vdash^{0}_{2} (\overline{r} \wedge \overline{t})}\,(\wedge)}{\vdash^{1}_{4} (p \vee q) \wedge (p \vee \overline{p}) \wedge (\overline{r} \wedge \overline{t})}\,(\wedge)$$

Here, it is possible to observe that for each step of the proof, we can directly read the fractional semantics value on the turnstile.

2.1. *Framing beliefs into Fractional Semantics for classical logic*

From Fractional Semantics we can do a different framework where beliefs are incorporated into fractional semantics for classical logic by introducing a set of axioms denoted as B. These axioms, representing the true beliefs of an agent, are treated as tautologies. The underlying philosophy is that an agent naturally considers their own beliefs to be true.

Beliefs in this context are treated as deductively closed, implying that any deduction made using these true beliefs is also considered true. This reflects the idea of an agent being deductively ideal. Integrating such beliefs into fractional semantics can lead to obtaining values greater than those typically permitted by fractional semantics alone.

The inspiration for this expansion comes from one of Makinson's methods, namely *pivotal-assumption consequence*, used to bridge the gap between classical and non-monotonic logic by adding background assumptions. However, the fractional semantics approach with added beliefs differs from *pivotal-assumption consequence* in two key aspects. Firstly, while Makinson used a classical two-valued semantics, fractional semantics operates within a multi-valued interpretation. Secondly, *pivotal-assumption consequence* assigns the value 0 if any axiom is not a proper axiom or belief, whereas fractional semantics can assign values greater than 0 when a top sequent is a tautology or a belief.

To incorporate beliefs into the system, they must be atomic; otherwise, they need to be decomposed. The definitions of $GS4_B$ and $\vdash_B$ are provided as follows:

Definition 2.5 ($GS4_B$). Let $\mathbb{B} = b_1, \ldots, b_n$ a set of non tautological, non contradictory and of arbitrary complexity formulas; let B be the set of sequents obtained from the decomposition of formulas in $\mathbb{B}$ and closed under cut; let $GS4$ be as defined earlier, then $GS4_B$ is the system where everything that is derived from $\mathbb{B}$ and from $GS4$ is true.

Definition 2.6 ($\vdash_{\boldsymbol{B}}$). If $\vdash$ is the closure relation of classical logic, then $\vdash_B$ is defined as the closure relation of $GS4_B$.

The system is not Post-complete because structurality and consistency are mutually exclusive properties in the axiomatic extension of classical logic: adding new axioms to the system is not possible to mantain structurality, *i.e.*, substitution is dropped. Makinson highlighted this in [15] without explicitly citing Post, even though the underlying reason is identical.

Theorem 2.7. *There is no supra-classical closure relation in the same language as classical $\vdash$ that is closed under substitution, except for $\vdash$ itself and the total relation* i.e. *the relation that relates every possible premises to every possible conclusion.*

and this applies also to this system.

Now, let's delve deeper into formalizing the system by defining the top sequent incorporating added beliefs.

Definition 2.8.

$top^b(\pi)$: represents the multiset of all and only top sequents of π introduced by a belief.

The reason for introducing this new type of top sequent stems from our desire, particularly in this context, to treat beliefs on par with identity axioms. This is because an agent invariably regards her own beliefs as true. The updated method for calculating the value of a sequent is:

$$[\![A]\!]_B = \frac{top^b(\pi) + top^1(\pi)}{top^b(\pi) + top^1(\pi) + top^0(\pi)}$$

It is also possible to see the same tree with multi valued system, adding a new rule:

$$\dfrac{}{\vdash^{1}_{1\,B} B}\,(\overline{b_i})$$

Example 2.9. For example let's consider this example where $B = p, q$

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\vdash^{1}_{1\,B} p, q}\,(b_1)}{\vdash^{1}_{1\,B} p \vee q}\,(\vee) \qquad \dfrac{\dfrac{}{\vdash^{1}_{1\,B} p, \overline{p}}\,(ax.)}{\vdash^{1}_{1\,B} p \vee \overline{p}}\,(\vee)}{\vdash^{2}_{2\,B} (p \vee q) \wedge (p \vee \overline{p})}\,(\wedge) \qquad \dfrac{\dfrac{}{\vdash^{0}_{1\,B} \overline{r}}\,(\overline{ax.}) \qquad \dfrac{}{\vdash^{0}_{1\,B} \overline{t}}\,(\overline{ax.})}{\vdash^{0}_{2\,B} (\overline{r} \wedge \overline{t})}\,(\wedge)}{\vdash^{2}_{4\,B} (p \vee q) \wedge (p \vee \overline{p}) \wedge (\overline{r} \wedge \overline{t})}\,(\wedge)$$

It is worth noting that if this sequent was considered in classical logic, any valuation would assign either the value 0 or 1. Something similar happens in Makinson pivotal assumption consequence, also if the belief set is the same that we have defined earlier, because a two valued logic is there considered.

3. *Strong cut elimination*

The last section pointed out that the agent is an ideal one and that they are aware of every deduction between beliefs. This means that the belief set is deductively closed: nothing that was not already in the set can be derived. In order to have a deductively closed belief set it is important that every combination of sentences, when it is possible, must be closed under cut and the new sentences obtained in this way will be added to the belief set.

In order to eliminate cut from $GS4_B$ the method is taken from [18], but it is simplified because of the nature of one-sided sequents. The method is the following:

1. let's consider a propositional formula $b_i \in B$ (B being the set of beliefs) and decompose it using the invertible rules;
2. the procedures gives identity and non-logical sequents. Remove the identity ones;
3. let's contract every sequent thus obtained;
4. let's consider two sequents Γ, p and $\Delta, \overline{p}$ and add the sequent Γ, Δ to the set of beliefs and let's contract the set thus obtained;
5. the procedure terminates;
6. finally, take the set closed under weakening.

To emphasize the importance of accounting for the fractional value of a formula incorporating beliefs, it is necessary to consider, as initial sequents, not only those obtained directly but also sequents derived via closure under cut. Let's illustrate this with the following example:

Example 3.1. It is easy to show why the step 4. is so important. Suppose that an agent has a new belief: $A = (\overline{p} \wedge (\overline{t} \vee q)) \vee (t \wedge (\overline{t} \vee q))$. The first thing to do in order to add that belief is to transform A in a conjunctive form: it is easy to show that it is equivalent to $\vdash (\overline{p} \vee t) \wedge (\overline{t} \vee q) \wedge (t \vee \overline{t} \vee q) \wedge (\overline{t} \vee q)$. Let's decompose it in a set of clauses: $\vdash \overline{p}, t$, $\vdash \overline{t}, q$, $\vdash t, \overline{t}, q$, $\vdash \overline{t}, q$ and remove one of the copies of $\vdash \overline{t}, q$ and the axiom $\vdash t, \overline{t}, q$. By the method presented earlier the agent has to add $(\overline{p} \vee t)$ and $(\overline{t} \vee q)$ to the system, but these beliefs are not cut free. To let them be cut free, it is necessary to close them under the cut.

$$\frac{\vdash \overline{p}, t \qquad \vdash \overline{t}, q}{\vdash \overline{p}, q}\ (cut)$$

From the last point of the method presented earlier, it is needed to add not only $\vdash \overline{p}, t$ and $\vdash \bar{t}, q$, but also $\vdash \overline{p}, q$. Let's see why: $[\![(\overline{p} \vee t) \wedge (\bar{t} \vee q)]\!]$ has value 1 if $B = \{(\overline{p}, t); (\bar{t}, q)\}$

$$\dfrac{\dfrac{\vdash^{1}_{1\,B} \overline{p}, t}{\vdash^{1}_{1\,B} \overline{p} \vee t}\,(\vee) \qquad \dfrac{\vdash^{1}_{1\,B} \bar{t}, q}{\vdash^{1}_{1\,B} \bar{t} \vee q}\,(\vee)}{\vdash^{2}_{2\,B} (\overline{p} \vee t) \wedge (\bar{t} \vee q)}\,(\wedge)$$

As it was showed, the cut is really important for a complete set of beliefs, but it is also necessary to see how the cut can be eliminated from the calculus.

3.1. *Elimination of cut*

The elimination of cut in presence of proper axioms was firstly proposed by Girard [6], as noted by Avron [1], upgrading the Gentzen's standard cut elimination algorithm. The procedure here proposed, *i.e.*, the decomposition of the formula, the add to the system and the cut of the formula to obtain all the derivations, owes a lot to the one presented in [18].

In the article, in fact, is proved that, for any cluster of extra-logical assumptions, there exists exactly one axiomatic extension of classical propositional logic that admits cut elimination. We can prove that Fractional value does not decrease in $GS4_B$ with relation to the addition of formulas:

Theorem 3.2. *For any multiset of atomic formulas* $\vdash_B \Gamma$ *and* $\vdash_B \Delta$, $[\![\bigvee \Gamma \vee \bigvee \Delta]\!]_B \geq [\![\bigvee \Gamma]\!]_B$.

Proof. To prove this is sufficient to consider a transformation of $\vdash_B$. In fact if $B = b_1, \dots b_n$, then $\vdash_B \Gamma$ is equal to $\vdash \Gamma, \overline{b_1}, \dots \overline{b_n}$, changing the kind of turnstile from the one introduced here to the classical one, as pointed out in [15][1]. Intuitively this is due to the fact that the sequent is true iff there is a disjunction between a letter and its negation (for example b_i and $\overline{b_i}$). From this fact it is possible to consider four cases:

- if $[\![\Gamma]\!]_B = [\![\Delta]\!]_B = 1$, than obviously $[\![\Gamma \vee \Delta]\!]_B = 1$ as well;
- if $[\![\Gamma]\!]_B = [\![\Gamma, \overline{b_1}, \dots, \overline{b_n}]\!] = 1$ and $[\![\Delta]\!]_B = 0$, then $[\![\Gamma \vee \Delta]\!]_B = 1$ as well;

[1] In the text the two sided version of this transformation was used, so $\vdash_B \Gamma$ becomes $b_1, \dots, b_n \vdash \Gamma$, but here because of the choice to use $GS4$ as main system, it is used the one-sided classically equivalent version $\vdash \Gamma, \overline{b_1}, \dots, \overline{b_n}$.

- if $[\![\Delta]\!]_B = [\![\Delta, \overline{b_1}, \ldots, \overline{b_n}]\!] = 1$ and $[\![\Gamma]\!]_B = 0$, then $[\![\Gamma \vee \Delta]\!]_B \geq [\![\Gamma]\!]_B$, whatever value assumes $[\![\Gamma]\!]_B$;
- if $[\![\Gamma]\!]_B = [\![\Delta]\!]_B = 0$, then $[\![\Gamma \vee \Delta]\!]_B \geq [\![\Gamma]\!]_B$. □

It is possible to generalize this result for any context:

Theorem 3.3. *For any context Γ and a formula A, such that A is not contradictory with the set B, $[\![\bigvee \Gamma \vee A]\!]_B \geq [\![\Gamma]\!]_B$.*

Proof. Let's prove it by induction on the complexity of the formula A.

Base case: Let's consider A atomic, then we have two cases:

$A \in B$: if $A \in B$, then $[\![\bigvee \Gamma, A]\!]_B = 1$ and $[\![\Gamma, A]\!]_B \geq [\![\Gamma]\!]_B$;
$A \notin B$: if $A \notin B$, then if $[\![\bigvee \Gamma]\!]_B = 1$, there is an atomic formula in Γ that is in the belief set, so also $[\![\bigvee \Gamma, A]\!]_B = 1$. If $[\![\bigvee \Gamma]\!]_B = 0$, $b_1, \ldots, b_n \notin \Gamma$ and then $[\![\bigvee \Gamma \vee A]\!]_B = 0$

Inductive step: Let's consider two cases:

$A \equiv p \wedge q$: by inductive hypothesis $[\![\bigvee \Gamma \vee p]\!]_B \geq [\![\Gamma]\!]_B$ and $[\![\bigvee \Gamma \vee q]\!]_B \geq [\![\Gamma]\!]_B$. If at least one between $[\![\bigvee \Gamma \vee p]\!]_B$ and $[\![\bigvee \Gamma \vee q]\!]_B$ is equal to 0, then $[\![\bigvee \Gamma]\!]_B = 0$ for inductive hypothesis and then $[\![\bigvee \Gamma \vee (p \wedge q)]\!]_B \geq [\![\Gamma]\!]_B$. The only remaining case is when $[\![\bigvee \Gamma \vee p]\!]_B = 1$ and $[\![\bigvee \Gamma \vee q]\!]_B = 1$:

$$\dfrac{\dfrac{\dfrac{}{\vdash^{1}_{1\,B} \Gamma, p}}{\vdash^{1}_{1\,B} \bigvee \Gamma \vee p}\,(\vee) \qquad \dfrac{\dfrac{}{\vdash^{1}_{1\,B} \Gamma, q}}{\vdash^{1}_{1\,B} \bigvee \Gamma \vee q}\,(\vee)}{\vdash^{2}_{2\,B} \bigvee \Gamma \vee (p \wedge q)}\,(\wedge)$$

Thus $[\![\bigvee \Gamma \vee (p \wedge q)]\!]_B \geq [\![\Gamma]\!]_B$.

$A \equiv p \vee q$: by Theorem 3.2. □

Theorem 3.4 (Strong cut elimination of $GS4_B$). *The cut rule is redundant when added to $GS4_B$.*

Proof. Girard was the first to notice that a different procedure could preserve cut elimination even in the presence of axioms [6,16]. The proof is as usual with double induction, the alorithm is similar to the one presented in [20]. □

The set of beliefs can be "completed" through cut or without that. This means that $GS4_B$ is a cut-free system, because it is an axiomatic extension of classical logic. By the way, the use of cut can alter the fractional semantics value as shown in [17]. Thanks to theorem 3.4 the algorithm presented in section 3 can be transformed in an algorithm without the presence of cut. As a corollary of the strong cut elimination it can be obtained:

Theorem 3.5 (Uniqueness of axiomatization in $GS4_B$). *For any cluster of axioms in the set of beliefs B the axiomatization is unique.*

Proof. See [18]. □

4. *Full Beliefs and Revisable Beliefs*

The formal model of beliefs introduced since here is a dichotomous system, an all-or-nothing structure, where a belief is either fully accepted or not at all. We have previously asserted that beliefs, within this framework, are deemed true as well as tautologies. However, we can refine this categorization further. In this section, we employ hyperreal numbers, as Hansson did [10], to distinguish between tautologies and beliefs, or more precisely, between Full Beliefs and Revisable Beliefs. These designations are arbitrary and simply signify that a Full Belief is one whose value is immutable, while a Revisable Belief is one that, though currently held as true, remains subject to revision in light of new evidence.

The reason why hyperreal numbers are interesting in this kind of settlement is twofold: on one hand, it is easy to distinguish between Revisable and Full Beliefs; on the other hand, hyperreal numbers do not alter the fractional final value, thereby validating all the proofs that we have made for $GS4_B$ also for this settlement. In fact, $1-\delta$ in $\mathbb{Q}$ is equal to 1, creating a bridge between $GS4_B$ and hyperreal numbers.

A Full Belief is characterized as a belief that remains impervious to revision under any circumstances; it is an assertion that an agent is unwilling to discard in any situation. Conversely, an agent may hold beliefs that are fully accepted, yet subject to revision in light of new evidence; these are termed Revisable Beliefs, in the sense that they are beliefs that, in presence of new evidences, can be revised, while a tautology can be regarded as a Full Belief, because it can't be revised also in presence of new evidences. For the sake of enhancing the generality of the system, we extend this classification beyond tautologies alone. To accomplish this, we adorn the turnstile with the expression $1-\delta$, where δ represents an infinitesimal quantity:

$$\frac{}{\mid\!\frac{1-\delta}{1}{}_{B}\; b_j}\;(b_j)$$

This notation implies that the belief b_j is one of the agent's beliefs, subject to possible revision in a subsequent moment. The symbol δ functions as a label derived during the proof, serving to keep track of the use of one or more propositions that may be revised in the presence of new evidences. This new notation doesn't change the proves of Cut and Weakening Admissibility in function [10]: $st(1-\delta)=1$, because of the fact that for the proofs we can use only the standard part of the hyperreal number. This means that from the point of view of proof theory nothing changed, but something changed in the expressivness of Fractional Semantics. The rule of conjunction still decorate the sequents in the same way:

$$\dfrac{\vdash^{m_1}_{\mathrm{n}_1} \Gamma, A \qquad \vdash^{m_2}_{\mathrm{n}_2} \Gamma, B}{\vdash^{m_1+m_2}_{\mathrm{n}_1+\mathrm{n}_2} \Gamma, A \wedge B}\ (\wedge)$$

the only difference is that sometimes we will have to add also infinitesimal numbers, for example:

$$\dfrac{\vdash^{1-\delta}_{1}{}_{B}\ p \qquad \vdash^{1-\gamma}_{1}{}_{B}\ q}{\vdash^{2-(\delta+\gamma)}_{2}{}_{B}\ p \wedge q}\ (\wedge)$$

Also, if the final value seems strange, it indicates that, according to Fractional Semantics, the value remains 1. This implies that the derivation is solely based on true assumptions at the moment of the derivation. On the other hand, we have two infinitesimals, suggesting that two of the assumptions are beliefs that can be discarded in the presence of new evidence. None of the beliefs used are Full Beliefs, so $p \wedge q$ is a proposition with a value of 1, thanks to the set B. The meaning of the value $2 - (\delta + \gamma)$ is that two of the leaves of the tree are beliefs that are possible to revise in the presence of new information. This means that, after revision, the fractional value could also assume a value of 0.5 or maybe also 0 if both of the beliefs once revised result as false. Our idea is that this value is a way to keep track of how many beliefs are not Full beliefs or tautologies into the derivation.

4.1. *Decomposition of a Revisable Belief*

To decompose a belief, the rules remain the same as before; we must decompose it and close under cut. Suppose we aim to incorporate the belief $\vdash q \wedge (r \vee \overline{s})$ into the system, but it is not a full belief. To achieve this, we need to decompose it:

$$\dfrac{\vdash q \qquad \dfrac{\vdash r, \overline{s}}{\vdash r \vee \overline{s}}\ (\vee)}{\vdash q \wedge (r \vee \overline{s})}\ (\wedge)$$

Now, to indicate that the original belief $\vdash q \wedge (r \vee \overline{s})$ was neither a Full Belief nor a Tautology, we adjust its value by adding the number $1 - \delta$ instead of 1. This adjustment accounts for the infinitesimal nature of δ, and its division by 2 ensures the preservation of infinitesimal characteristics.

$$\cfrac{\vdash^{1-\delta}_{1}{}_{B}\, q \qquad \cfrac{\vdash^{1-\gamma}_{1}{}_{B}\, r, \overline{s}}{\vdash^{1-\gamma}_{1}{}_{B}\, r \vee \overline{s}}\,(\vee)}{\vdash^{1-(\delta+\gamma)}_{2}{}_{B}\, q \wedge (r \vee \overline{s})}\,(\wedge)$$

In the event that either $\vdash\ q$ or $\vdash\ r, \overline{s}$ is employed in a derivation, we explicitly denote this value in the sequent derivation. For instance:

$$\cfrac{\cfrac{\cfrac{\vdash^{1}_{1}{}_{B}\, p, \overline{p}}{\vdash^{1}_{1}{}_{B}\, p \vee \overline{p}}\,(\vee) \qquad \vdash^{1-\delta}_{1}{}_{B}\, q}{\vdash^{2-\delta}_{2}{}_{B}\, (p \vee \overline{p}) \wedge q}\,(\wedge) \qquad \cfrac{\vdash^{0}_{1}{}_{B}\, \overline{p}, q}{\vdash^{0}_{1}{}_{B}\, \overline{p} \vee q}\,(\vee)}{\vdash^{2-\delta}_{3}{}_{B}\, (p \vee \overline{p}) \wedge q \wedge (\overline{p} \vee q)}\,(\wedge)$$

This implies that, even without knowing the initial values of the leaves, we can still make observations about the value $2-\delta/3$: the standard part of the derivation is the Fractional Semantics value in $GS4_B$ is $2/3$ and we can observe that there is only one infinitesimal number, indicating that only one of the initial beliefs is a Revisable Belief. The portion that is neither a Full Belief nor a Revisable Belief is then $1/3$, representing what remains between $2/3$ and 1.

5. *Conclusions*

The current endeavor to unite Full Beliefs, Revisable Beliefs, and tautologies represents an initial stride towards establishing a connection between Fractional Semantics and Probability. Fractional Semantics emerges as a powerful instrument for delineating the intricacies of a derivation, offering valuable insights into the dynamic evolution of belief values and the interplay between Revisable Beliefs and Full Beliefs throughout the proof. This amalgamation serves as a foundational framework, setting the stage for a more comprehensive exploration of the relationship between Fractional Semantics and Probability.

The forthcoming phase of our research will delve into elucidating the intricate links between Fractional Semantics and Belief Revision. This constitutes another pivotal facet that underscores the significance of the introduced system. The versatility of our system, embracing both the stability of Full Beliefs and the adaptability of Revisable Beliefs, positions it as an invaluable tool for delving into the nuances of belief dynamics and their evolution over the course of iterative revisions. By bridging the gap between Fractional Semantics and Belief Revision, we aim to provide a more holistic understanding of the nuanced interplay between formal semantics and the adaptive nature of belief systems.

References

[1] A. Avron, *Gentzen-type systems, resolution and tableaux*, J. Automat. Reason. **10** (1993), 265–281.

[2] M. Bizzarri, *A solution to the lottery paradox through fractional semantics*, 5th Pisa Colloquium (2023).

[3] M. Bizzarri, *Framing beliefs into fractional semantics for classical logic*, 5th SILFS Postgraduate Conference Proceeding (2024).

[4] R. Foley, *The Epistemology of Belief and the Epistemology of Degrees of Belief*, Amer. Philos. Quart. Monogr. Ser. **29** (1992), 111–124.

[5] P. Gärdenfors and D. Makinson, *Revisions of Knowledge Systems Using Epistemic Entrenchment*, In: "Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge", TARK '88, Morgan Kaufmann Publishers Inc., Pacific Grove, California, 1988, 83–95.

[6] J.-Y. Girard, "Proof and Types", Cambridge university press, 1989.

[7] A. Grove, *Two modellings for theory change*, J. Philos. Logic **17** (1988), 157–170.

[8] J. Y. Halpern, "Reasoning about Uncertainty", MIT Press, Cambridge, MA, USA, 2003.

[9] S. O. Hansson, *A basis for* AGM *revision in bayesian probability revision*, J. Philos. Logic **52** (2023), 1535–1559.

[10] S. O. Hansson, *Revising probabilities and full beliefs*, J. Philos. Logic **49** (2020), 1005–1039.

[11] J. Hawthorne, *The Lockean thesis and the logic of Belief*, In: "Degrees of Belief", F. Huber and C. Schmidt-Petri (eds.), Synthese Library: Springer, 2009, 49–74.

[12] H. Dominic, *A minimal classical sequent calculus free of structural rules*, 2005.

[13] S. C. Kleene, "Mathematical Logic", John Wiley & Sons, 1967.

[14] H. Leitgeb, "The Stability of Belief: How Rational Belief Coheres with Probability", Oxford University Press, 2017.

[15] D. Makinson, "Bridges from Classical to Nonmonotonic Logic", Lightning Source, Milton Keynes, 2005.

[16] H. Pfeifer and J.-Y. Girard, *Proof theory and logical complexity*, J. Symb. Log. **54** (1989), 1493 p.

[17] M. Piazza and G. Pulcini, *Fractional semantics for classical logic*, Rev. Symb. Log. **13** (2020), 810–828.

[18] M. Piazza and G. Pulcini, *Uniqueness of axiomatic extensions of cut-free classical propositional logic*, Log. J. IGPL **24** (2016), 708–718.

[19] M. Piazza, G. Pulcini and M. Tesi, *Fractional-valued modal logic*, Rev. Symb. Log. (2021), 1–22.

[20] M. Piazza and M. Tesi, *Analyticity with extra-logical information*, J. Logic Comput. (2024).

[21] R. Smullyan, "First-Order Logic", Courier corporation, 1995.

[22] A. S. Troelstra and H. Schwichtenberg, "Basic Proof Theory", 2 edition, Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, 2000.

Matteo Bizzarri

From expert testimony to lay belief: a Bayesian view

1. *Introduction*

Modern societies are crucially dependent on the opinion of experts. The asymmetric relationship between experts and non-experts poses many socially impactful problems that have been subjected to the lens of different disciplines, including philosophy of science, social epistemology, argumentation theory, cognitive science, and psychology. In this paper, we focus on one fundamental problem that lies at the intersection of these approaches: how lay reasoners should update their beliefs in some hypothesis or claim H given that some expert asserts that H. We will approach this question from the perspective of Bayesian epistemology, following a recent trend that applies Bayesian models to represent, clarify, and manipulate traditional concepts in all of the above-mentioned areas [3,9,21,29,33,37]. More precisely, we shall treat the assertion of H by the expert as a case of testimony, and study how a Bayesian layman should update his probabilistic beliefs in H given this evidence. Bayesian confirmation theory [8, 13] will be helpful in modeling the notion of expert reliability and how it affects laymen's reasoning.

We proceed as follows. In section 2 we introduce the elements of the Bayesian approach to uncertain reasoning and confirmation of hypotheses. In section 3, we model expert opinion as evidence from testimony, developing a Bayesian model of both the experts' reliability and the evidentiary impact of their testimony in terms of Bayesian confirmation measures. In section 4, we show how our model impacts the discussion concerning expert opinion both in social epistemology and argumentation theory. As for the former, we show how the idea of "epistemic deference" amounts to ignoring the fallibility of real experts and fails as a general strategy of belief updating in the face of expert opinion. As for the latter, we elaborate on the recent discussion of *ab auctoritate* reasoning to clarify and evaluate this much-disputed argumentative strategy. By adopting a model of expert reliability based on Jeffrey conditionalization, we argue for a new characterization of *ab auctoritate*, more faithful to the complex epistemic interplay between laymen and experts. A short conclusion and some ideas for future research are offered in section 5.

2. *Bayesian reasoning, confirmation, and evidence*

In the following, we will treat both laymen and experts as ideal epistemic agents in the Bayesian sense. In this section, we offer a quick overview of the Bayesian framework, which will be applied in the next sections to the problem of expertise and testimony.

Updating beliefs on evidence. If Ann is a Bayesian agent, her beliefs about a given hypothesis H (such as "It will rain tomorrow") are represented by the (subjective) probability $p(H)$ she assigns to H. When Ann receives some piece of evidence E relevant to H (like "Tonight the sky is cloudy"), Ann will change her belief in H by moving from $p(H)$ to the conditional probability $p(H|E)$ of H given E. The updating rule is governed by the Bayes theorem:

$$p(H|E) = \frac{p(E|H)p(H)}{p(E)} \tag{2.1}$$

where $p(H)$ is often called the "prior" probability of H and $p(H|E)$ its "posterior" probability (viz. before and after receiving evidence E). Bayes rule makes clear that the strength of Ann's posterior belief in H should be proportional both to that of her prior belief in H and to the so-called *likelihood* of H, *i.e.*, the probability $p(E|H)$ expressing how well H "explains" E or, better, how much E is expected in the light of H.[1]

For most applications, it is convenient to express the denominator of the Bayes formula, *i.e.*, the probability of the evidence, in its "unpacked" form using the rule of total probability. The resulting updating rule, which is equivalent to eq. (2.1) above, will look like this:

$$p(H|E) = \frac{p(E|H)p(H)}{p(E|H)p(H) + p(E|\neg H)p(\neg H)} \tag{2.2}$$

This new formulation of Bayes theorem makes clear that $p(H|E)$ must be a function of three different probabilities: the prior $p(H)$ of the hypothesis, its likelihood $p(E|H)$, and the likelihood of its negation, *i.e.*, $p(E|\neg H)$. Intuitively, H is the more probable in the light of E the more H was probable before observing E and the more E was expected given H instead of $\neg H$.

[1] For an introduction to Bayesian reasoning see [20]; see also [23] and [33] for more advanced discussions.

To make things more concrete, suppose that Ann is going to take a COVID-19 test in order to see whether she is ill (H) or not ($\neg H$). Before taking the test, she may believe she has a 5% probability of being ill, perhaps on the basis of her knowledge about the prevalence of the disease in her area. This fixes her prior probability $p(H)$ to 0.05 (and her prior probability of *not* being ill to $p(\neg H) = 1 - p(H) = 0.95$). Consulting the test's leaflet, Ann can also find the two likelihoods required by Bayes theorem. The "sensitivity" of the test, or, in other words, its true positive rate, expresses the probability $p(E|H)$ that the test will turn out positive (E) if Ann were actually ill; let's assume that $p(E|H) = 0.7$. The "specificity" $1 - p(E|\neg H)$ of the test is instead related to its false positive rate, expressing the probability that Ann gets a positive result even if she is actually not ill; let's assume that the specificity is 90%, and hence $p(E|\neg H) = 0.1$. Putting these numbers into eq. (2.2) gives us:

$$\begin{aligned} p(H|E) &= \frac{p(E|H)p(H)}{p(E|H)p(H) + p(E|\neg H)p(\neg H)} \\ &= \frac{0.7 \times 0.05}{(0.7 \times 0.05) + (0.1 \times 0.95)} \simeq 0.27 \end{aligned}$$

Thus, Ann has a 27% probability of having COVID-19 given that the test result was positive. Note that this figure is significantly lower than 50%, meaning that it is still more probable that Ann is fine than she is ill. The reason is that we assumed that being ill was quite improbable (5%) in the first place; still, such probability has much increased (more than five times) after observing the test result. This increase in probability signifies how much the evidence given by the positive result "supports" or "confirms" the hypothesis of illness. Bayesian confirmation theory aims at making this intuition precise.

Confirmation as evidential support. Another interesting formulation of Bayes theorem is obtained expressing the relevant probabilities, following the gambling use, in terms of "odds". In statistical jargon, the odds of a hypothesis H refers to the ratio $o(H) = \frac{p(H)}{p(\neg H)}$ of the probability of H to that of its negation. Re-writing Bayes theorem in terms of these ratios leads to its "odds" form:

$$\frac{p(H|E)}{p(\neg H|E)} = \frac{p(E|H)}{p(E|\neg H)} \times \frac{p(H)}{p(\neg H)} \tag{2.3}$$

In this new form, the theorem tells us that the ratio of the posterior probabilities of H and its negation is given by the ratio of the prior probabilities multiplied by the so-called "likelihood ratio", *i.e.*, the ratio of the likelihoods of H and of its negation. Adopting the odds notation and writing "$LR(H,E)$" for the likelihood ratio, we can re-write eq. (2.3) as follows:

$$o(H|E) = LR(H,E) \times o(H) \tag{2.4}$$

That is to say, the posterior odds of H given E are just its prior odds multiplied by the likelihood ratio.[2]

The advantage of this odds formulation is that it makes very clear the role of the evidence E in changing Ann's belief in H. Let's consider again our previous example relative to Ann taking a COVID-19 test. Initially, the odds of Ann in favor of H (the hypothesis of being ill) are very low: $o(H) = \frac{0.05}{0.95} \simeq 0.053$. However, when Ann takes into account the positive test result, she can calculate a likelihood ratio of $LR(H,E) = \frac{p(E|H)}{p(E|\neg H)} = \frac{0.7}{0.1} = 7$. This number tells Ann how much the evidence received by the test speaks in favor of H rather than of its negation: multiplying her prior odds $o(H)$ by this factor gives her posterior odds in favor of H:

$$o(H|E) = LR(H,E) \times o(H) = 7 \times 0.053 \simeq 0.37$$

Thus, H is now much more likely than before, indeed seven times more.[3]

More generally, the likelihood ratio LR is a possible measure of how strongly a piece of evidence E supports or undermines a given hypothesis H, *i.e.*, of how much H is confirmed or disconfirmed by the evidence.[4] In Bayesian confirmation theory, we say that H is confirmed by E iff the probability of H is increased in the light of E: *i.e.*, $p(H|E) > p(H)$; that H is disconfirmed by E iff the probability of H is decreased in the light of E: *i.e.*, $p(H|E) < p(H)$; and that E is neutral for H iff the probability of H is left untouched by learning E: *i.e.*, $p(H|E) = p(H)$. It is now not difficult to prove that $LR(H|E) > 1$ iff E confirms H; that $LR(H|E) < 1$ iff E disconfirms H; and that $LR(H|E) = 1$ iff E is neutral for H. In the practice of many scientific disciplines, for instance medicine, the likelihood ratio is routinely employed (often under the name of "Bayes factor") as a measure of how well some evidence (*e.g.*, the result of a medical test)

[2] Note that some authors – like [23, 21] and [3, 17] – call "likelihood ratio" the reciprocal of $LR(H,E)$, *i.e.*, $\frac{p(E|\neg H)}{p(E|H)}$. We follow here the standard terminology in contemporary Bayesian confirmation theory [8, 33].

[3] If one has the odds o in favor of H, one can calculate its probability p as follows: $p = \frac{o}{o+1}$. As one can check, with this formula the just calculated odds of 0.37 gives us the 27% probability of section 2.

[4] Note that there are many, non-equivalent, confirmation measures, and many arguments in favor or against each of them [8, 14, 15, 33].

supports or undermines a given hypothesis (*e.g.*, a diagnosis). In Section 3 we shall return to the role played by likelihood ratios or Bayes factors in the confirmation of hypotheses.

Updating beliefs on uncertain evidence. Above, we employed Bayes theorem to say how Ann has to rationally update her belief in the face of incoming evidence. In other words, we assumed that Ann had a certain degree of belief $p(H)$ in the truth of H at some moment t_0, after which she receives some evidence E. At the new time t_1, Bayesian conditioning prescribes that Ann's degree of belief in the truth of H changes from her old credence $p(H)$ to her new credence $p'(H)$, which must be equal to $p(H|E)$ as given by eq. (2.1). Crucially, we assumed here that E is veridical and certain, so that, after receiving E, $p'(E) = 1$. This means, for instance, that there are no doubts in the reading of the COVID-19 test result, which is certainly positive in our example above.

Now suppose that just before the result is displayed on the test, a power failure occurs in Ann's room. In the dark, Ann cannot be completely confident about the outcome appearing on the test. Compared with the previous case, where Ann knew with certainty whether the test was positive or not, there is now an added dimension of uncertainty concerning the interpretation of evidence. This kind of uncertainty cannot be addressed by standard Bayesian conditioning. Instead, Ann's belief updating will have to follow "Jeffrey Conditioning" [24, 25]:

$$p'(H) = p(H|E)p'(E) + p(H|\neg E)p'(\neg E) \tag{2.5}$$

In words, the new probability Ann assigns to H is the sum of her old probabilities in H given E and in H given $\neg E$, respectively weighted by the (new) probability that E is actually true.[5]

For example, suppose Ann is fairly confident that she sees a double line (*i.e.*, a positive result). Let us set her subjective probability to $p'(E) = 0.8$. Her belief updating on this uncertain evidence will then be as follows:

$$\begin{aligned} p'(H) &= p(H|E)p'(E) + p(H|\neg E)p'(\neg E) \\ &= 0.27 \times 0.8 + 0.017 \times 0.2 = 0.22 \end{aligned}$$

Not surprisingly, Ann's degree of belief in H is now lower than in the previous case (22% as opposed to 27%), due to the uncertainty of the evidence.

[5] Note that if Ann is certain of her evidence, *i.e.*, $p'(E) = 1$ and hence $p'(\neg E) = 0$, then Jeffrey conditionalization given by eq. (2.5) immediately reduces to standard Bayesian conditionalization as in eq. (2.1).

This concludes our survey of the Bayesian framework employed in this paper, which is applied to the problem of expert testimony in the next section.

3. *A Bayesian model of expert testimony and reliability*

When Bob, a layman, listens to the opinion of some recognized expert, Bob's beliefs will typically change in one way or another. Such change is the result of a complex interaction between the two, which usually involves both epistemic and non-epistemic aspects (*e.g.*, trust), that are being studied across different disciplines (including social epistemology and philosophy of science, social and cognitive psychology, political science, and so on). Here, we shall focus on the purely epistemic dimension of the layman-expert relationship.

More precisely, we shall assume that Bob is a Bayesian agent and study how his (probabilistic) beliefs in some hypothesis H should change when some recognized expert testifies that H is true. In general, we can think of a (scientific) expert as a particular kind of witness, who has privileged epistemic access to a particular state of the world that falls within her domain of expertise. Intuitively, Bob's credences will depend, at least in part, on the *reliability* of the expert. The more the expert is reliable, the more the expert testimony should impact on Bob's belief in H. In this section, we apply the framework introduced in Section 2 to analyze these issues.[6]

Expert testimony as evidence. Suppose that Bob, who has some prior beliefs concerning H, learns that an expert has testified that H is true. The expert's opinion is the evidence on which Bob updates his beliefs. To reflect this in the notation, we shall write "E_H" for "expert E testifies that H is true". For the sake of simplicity, we shall assume that the expert's testimony is "categorical": *i.e.*, it corresponds to a clear-cut affirmation or negation of the hypothesis under evaluation, without considering degrees of certainty. In other words, we only consider cases where the expert says, for instance, that it will rain (for sure) tomorrow (E_H), but not cases where the expert says that it is likely that it will rain tomorrow, or that it will rain with, say, 80% probability. Thus, we shall focus on Bob's posterior assessment $p(H|E_H)$ of the probability of H given the expert testimony in favor of H.

[6] A problem we are not going to discuss here, and that has attracted much attention in the recent philosophy of science literature, starting at least from [3], is whether multiple pieces of evidence coming from independent sources of information confirm a hypothesis more than a less "varied" *corpus* of evidence (the so-called Variety of Evidence Thesis). For Bayesian analyses of this problem see, *e.g.*, [5, 6, 22], and references provided there.

From eq. (2.1) and eq. (2.2), we know how Bob should update his beliefs in H:

$$\begin{aligned} p(H|E_H) &= \frac{p(E_H|H)p(H)}{p(E_H)} \\ &= \frac{p(E_H|H)p(H)}{p(E_H|H)p(H) + p(E_H|\neg H)p(\neg H)} \end{aligned} \tag{3.1}$$

In words, Bob's posterior belief in H, given the expert's testimony that H, is a function of his previous belief in H, represented by the prior $p(H)$, and of the two likelihoods associated to the testimony: the probability $p(E_H|H)$ that the expert (correctly) reports H when H is true, and the probability $p(E_H|\neg H)$ that the expert (mistakenly) reports H when H is false.

Interestingly, this way of construing the testimony of a well-informed witness as a piece of evidence in hypotheses evaluation is common in the legal domain [1, 16]. For example, in a murder trial, a judge might be unsure whether the suspect was at the crime scene when the murder happened (H). By asking a witness, the judge might gain useful insights and update his confidence in H accordingly. A positive report by the witness will likely increase the judge's subjective probability of the suspect being at the crime scene, given that the witness is reliable (*e.g.*, he was a bystander and does not have any interest in reporting the false). Conversely, a negative report by a reliable witness will decrease the judge's confidence in H. The more reliable the witness, the stronger the evidential support of his testimonial report. As we shall see, this way of modeling testimony is useful also in studying the layman-expert relation.

Expert reliability as confirmation. As we have noted, the strength of expert testimony will depend on the expert's reliability. But what does it mean for an expert witness to be reliable from a Bayesian perspective? To express expert reliability in our framework, we shall assume what we may call the "diagnostic-test model" of expert witnesses *cf.* [3, Chapter 3]. In other words, we shall assume that an expert, like the COVID-19 test considered in Section 2, also has rates of true and false positives, expressed, respectively, by the likelihoods $p(E_H|H)$ and $p(E_H|\neg H)$. In the case of medical tests, such probabilities are readily available in the test leaflet; in the case of experts, it is instead difficult to quantify precisely their "sensitivity" and "specificity". Still, it is a useful idealization to construe the reliability of an expert as a function of those likelihoods.[7]

[7] Moreover, one can argue that there are (fallible) indicators, discussed at length in the expertise literature [7, 18, 30], that may help in evaluating at least approximately the reliability of experts.

Accordingly, we shall define the reliability $Rel_E(H)$ of expert E relative to hypothesis H as the relevant likelihood ratio $LR(H, E_H)$:[8]

$$Rel_E(H) = LR(H, E_H) = \frac{p(E_H|H)}{p(E_H|\neg H)} \tag{3.2}$$

This is to say that the expert testimony is more reliable the more likely it is that the expert testifies that H is true when H is in fact true, and the less likely it is that the expert testifies that H is true when H is instead false. Note that this amounts to expressing the expert's reliability in terms of the confirmation that the expert report E_H confers on H [13]. This becomes clear if one applies Bayes theorem in its odds form to the case at hand, *i.e.*, eq. (2.4):

$$o(H|E_H) = Rel_E(H) \times o(H) \tag{3.3}$$

This is to say that, after receiving the expert report E_H, Bob will revise his prior odds by multiplying them by the reliability of the expert. This means that the more reliable the expert is, the more his report will confirm H. In general, we may say that the expert is "reliable" (relative to H) when $Rel_E(H)$ is greater than 1; and "unreliable" if $Rel_E(H)$ is smaller than or equal to 1, since in this latter case his report will not increase, but possibly decrease, the probability that H is true.[9]

Again, the legal domain provides a useful illustration of this idea. Consider, as a toy example, a murder case: during a crowded football match, a supporter gets shot and dies. The murder weapon is readily found, but nobody has seen the shooter. A man, Mr. Guy Hapless, is seen walking briskly towards the exit. The authorities stop him and get his fingerprints. The judge assigned to the case then asks a forensic dactyloscopy expert to compare them with those found on the gun. The expert reports that the two fingerprints match. How should a Bayesian judge evaluate such evidence?

[8] See also [3, 9, 13]. Note that $Rel_E(H)$ "measures the reliability of the witness *with respect to the report in question* [here, H] and not the reliability of the witness *tout court*", as [3, 17] put it. The latter reliability will depend on the expert's (average) performance on several different hypotheses in some field; reliability $Rel_E(H)$ as defined here only measures the expert's performance on a single hypothesis H. Note that much recent literature follows [3] in modeling reliability in a different way, using Bayesian networks; a minimal network will comprise three nodes, one for the hypothesis, one for the witness report (our E_H), and one for the reliability of the witness, which can be inferred *a posteriori* by manipulating the other nodes *cf.* [29, Chapter 10].

[9] Bovens and Hartmann [3], who define expert reliability as $1 - \frac{1}{LR(H,E_H)}$, consider as "fully unreliable" an expert (or, more generally, a witness) for whom $Rel_E(H) = 1$; here, we also consider the case of "misleading" witnesses, for which $Rel_E(H) < 1$, whose testimony actually disconfirms the hypothesis under consideration.

Suppose that, initially, the judge has no specific reason to suspect that "Mr. Hapless is guilty" (H), besides the fact that he was attending the match. Moreover, let's suppose that the judge is (rightly or wrongly) sure that the fingerprints found on the gun will certainly be the murderer's. This means that the prior probability of H is $p(H) = \frac{1}{10000}$ and its prior odds $\frac{1}{9999}$, both very low. The expert's report that Mr. Hapless' fingerprints match those found on the gun will of course confirm H. How much will depend on the expert's reliability: let's assume that our expert is highly reliable, detecting true positives in 92.5% of the cases, and reporting a false positive only 0.1% of the times.[10] Then, our expert's reliability $Rel_E(H)$ will be equal to 925 and, applying eq. (3.3), we obtain:

$$o(H|E_H) = Rel_E(H) \times o(H) = 925 \times \frac{1}{9999} \simeq 0.092$$

which corresponds to a posterior probability of H approximately equal to 8%. Not surprisingly, the new probability $p(H|E_H)$ is significantly higher than the prior (roughly 800 times so), since the expert testimony has strongly confirmed the hypothesis of guilt. Given the low initial probability, however, such strong confirmation is not enough to think that Mr. Hapless, on this only evidence, is more likely guilty than not.

In this connection, let us note that some conventions have been established in the literature on Bayes factors (following seminal work by Turing and Good) to qualitatively assess the "weight of evidence" in favor of H provided by different likelihood ratios. For instance, in our example above the weight of evidence provided by the expert is well above the highest threshold of "decisive" evidence according to table 1. Interestingly, such literature may provide guidance in developing conventional expertise thresholds for any given scientific domain in light of its epistemic status.

TABLE 1 Interpreation of likelihood ratios LR as a measure of the weight of evidence according to [26] and [27].

LR	Weight of evidence	
	[26]	[27]
$1 - 3.2$	Barely worth mentioning	Barely worth mentioning
$3.2 - 10$	Substantial	Substantial
$10 - 32$	Strong	Strong
$32 - 100$	Very strong	Strong
> 100	Decisive	Decisive

[10] Expert performance in recognizing matching fingerprints has been the object of several studies [28,34,35]. For the sake of our example, we have selected the results that show higher expert reliability, as reported in [35].

4. *Bayesian rationality, epistemic deference, and arguments from authority*

In this section, we show how the tools from Bayesian epistemology that we have presented so far allow us to shed light on two much-debated topics in social epistemology and argumentation theory: epistemic deference and *ab auctoritate* arguments.

Epistemic deference vs. Bayesian rationality. Social epistemology studies the social aspects of knowledge production and dissemination, including the role of testimony, trust, and communication in the pursuit of knowledge by individuals and groups [19]. A fundamental question in this area is how one should be guided by other people's beliefs. One strategy discussed in the literature is *complete epistemic deference*, which is applied to interactions with experts, *e.g.*, [12]. This strategy implies that, when Bob receives the expert opinion E_H, he will have to revise his beliefs by adjusting them to the expert's. According to [12, 480], this means that if $p_{exp}(H)$ is the probability that expert E assigns to H, then Bob should simply adopt such probability as his own, so that $p_{Bob}(H) = p_{exp}(H)$. Since we are only considering categorical opinions from experts, this would imply that, if the expert testifies that H, then Bob should plainly accept H as certain, *i.e.*, $p_{Bob}(H|E_H) = 1$.

Epistemic deference is of course an extreme updating strategy in the face of expert opinion. It amounts to treating the expert as a fully veridical and trustworthy oracle. From a Bayesian standpoint, this strategy cannot be justified in general; the only case where it could apply is when two highly idealized assumptions are made:

1. experts are completely reliable;
2. laypeople's prior beliefs on H should not be taken into account.

As to (1), in our framework a completely reliable expert would be one who cannot make mistakes, *i.e.*, for which $p(E_H|\neg H) = 0$; this would make $Rel_E(H)$ meaningless, or better tending to infinity as $p(E_H|\neg H)$ approaches 0. Since we are interested in modeling real, fallible experts, we exclude this case from consideration. As for (2), priors play of course a crucial role in the Bayesian treatment of how our layman Bob reacts to the expert testimony (*cf.* eq. (3.3)). Here, a natural objection to such treatment is the following: it is sensible to exclude the layman's prior beliefs from the picture, since the expert holds (by definition) more accurate beliefs, hence the layman can just rely on the expert and needs not consider his own priors when updating on an expert report.

While intuitively convincing, we think one should rebut this objection, as our fictitious murder example shows. In that case, the Bayesian judge is not "deferent" to the fingerprint expert at all, at least not in the sense suggested by [12]. For sure, the judge duly takes into account the expert report as a crucial piece of evidence, without doubting it. This is as it should be, since while the judge is an expert in the legal domain, he should be regarded as a layman when it comes to recognizing matching fingerprints – and this is why the opinion of the forensic dactiloscopy expert is required.[11] However, the judge does not conclude that Mr. Hapless is guilty, as his prior for such a hypothesis is very low. This would happen if, as per (2) above, the judge ignores his own priors, meaning that $p(H) = p(\neg H) = 0.5$. Then, since $o(H) = 1$, eq. (3.3) would correctly reduce to $o(H|E_H) = Rel_E(H)$, so that the expert's opinion fully determines the posterior beliefs of the judge. However, in most cases, this would amount to committing the so-called prosecutor's fallacy, an instance of the "base rate fallacy", where the epistemic agent tends to conflate the posterior probability for the likelihood $p(E|H)$ [1, 16, 17].

The importance of the layman's priors in updating on expert evidence highlights an important distinction between two concepts: the "weight of evidence" in favor of H and the "acceptance" of H. The former is provided by the expert's report and appropriately measured, as we argued, by his reliability $Rel_E(H)$; the latter crucially depends, in addition, also on the layman's priors. The idea of epistemic deference tends to obscure such distinction and should be resisted accordingly.

Arguments from authority and layman testimony. Argumentation theory studies the social practice of "giving and asking for reasons" [4], focusing on reasonable and fallacious uses of arguments in dialogical contexts [11, 22]. An important case is that of arguments from authority (*ab auctoritate*). This strategy is employed whenever an arguer appeals to the opinion of an authority to support her standpoint in a critical discussion. The kind of authority we are interested in here is an epistemic one – and therefore this argument is also known as *appeal to expert opinion* [36]. For example, Bob and Carl are discussing whether it is true that food that has fallen on the floor within 5 seconds will still be safe to eat (H). Carl is confident that H while Bob is convinced of the opposite. One week before the conversation, Carl overheard on the TV an expert reporting that H (*i.e.*, Carl came to know the evidence E_H). To persuade Bob, Carl appeals to the authority of that expert claiming that H. This means that Bob has received some evidence from Carl's testimony, which we will denote by

[11] More generally, the epistemic roles of "layman" and "expert" are always domain-dependent: the same person can be an expert in one domain and a layman in another (see [18]).

L_{E_H} (where L is for "layman"). How is Bob supposed to change his mind in light of this evidence?

In argumentation theory, this scenario has been studied mostly from a procedural perspective (rather than an epistemic one): *i.e.*, what is reasonable for a discussant to ask in such a scenario, rather than how they should update their beliefs. Walton [36] has come up with a set of "critical questions" that a lay discussant is always entitled to ask. Their function is to momentarily shift the burden of proof on the proponent of the argument, by identifying its potential weak spots.[12] While these questions carry a useful heuristic value, they do not inherently offer criteria for evaluation of the answers. As we argue, a Bayesian approach provides valuable insights into the assessment of arguments from authority.

While some authors have explored this idea [9,21], what is still needed is a framework that specifically takes into account the significance of second-order testimony, viz. Bob's assessment of the evidence provided by Carl's report of an expert testimony. This, we submit, is the crucial aspect of interest in the analysis of arguments from authority, which distinguishes such case from that of "standard" expert testimony. In fact, this scenario adds a further dimension of uncertainty, compared to the fingerprint example in the previous paragraph. Not only is the link between H and E_H uncertain because of the fallibility of the expert providing the testimony; but, crucially, this evidence itself is also uncertain because it is not gained first-hand, but reported by a (lay) witness. Many things could go wrong and lead Carl to misreport what he heard – for example: Carl's bad faith, confirmation bias, motivated reasoning, failure of memory, hearing impairment, lack of understanding of the general context where the testimony was provided, *etc.*

To deal with this additional layer of uncertainty, we suggest proceeding as in Ann's example: like the sudden darkness of the room introduced an element of disturbance and opacity in the reading of the COVID-19 test, here Bob will have to deal with Carl's opaquely reported evidence. This is where Jeffrey Conditionalization comes back into the picture. Building on eq. (2.5), we can represent belief updating on a layman's report of expert testimony as follows:

$$p'(H) = p(H|L_{E_H}) = p(H|E_H)p'(E_H) + p(H|\neg E_H)p'(\neg E_H) \quad (4.1)$$

In words, Bob's new probability for H, given Carl's report L_{E_H} that an expert testified that H is true, will depend on the posterior probability of H

[12] They are: 1. How credible is expert *E* as an expert source?; 2. Is *E* an expert in the field *F* that assertion *A* is in?; 3. What did *E* assert that implies *A*?; 4. Is *E* personally reliable as a source?; 5. Is *A* consistent with what other experts assert?; 6. Is *E*'s assertion based on evidence?

being true given that the expert really testified that H – $p(H|E_H)$ – and, at the same time, on the probability of the hypothesis still being true in the light of the absence of the expert testimony – $p(H|\neg E_H)$.[13] Each of these values will have to be adjusted, respectively, for the probability that layman Carl reports the expert testimony correctly – $p'(E_H)$ – or incorrectly – $p'(\neg E_H)$. To assess such probabilities, Bob cannot rely on some objective assessment of Carl's reliability, contrary to what the judge could do, in our previous example, with the fingerprint expert. Indeed, one could see the difference between expert and layman witnesses precisely in this: that only for the former it is possible to estimate, at least approximately, the corresponding reliability. For the latter, Bob can only rely on his own subjective assessment of the reliability of Carl correctly reporting the testimony he has witnessed. It remains to be explored how Walton's critical questions can be useful for refining such an assessment.

5. *Conclusion*

In this paper, we addressed the epistemic problem of expertise through the lenses of Bayesian epistemology. First, we showed how a layman should update his beliefs on the basis of that particular kind of evidence provided by expert testimony, and how experts' reliability can be construed as the confirmation provided to the relevant hypothesis at issue. Then, we contrasted our Bayesian model with the strategy of complete epistemic deference as discussed in social epistemology, and argued that it can also shed light on arguments from authority. Future work on these lines can move in different but complementary directions. As to social epistemology, the Bayesian approach seems promising when applied to problems of peer-expert disagreement [31]. In argumentation theory, Bayesian insights can help in bridging the gap between procedural and epistemic approaches to reasonable dialogues [2]. Finally, the Bayesian standard offers a benchmark against which both experts' and laymen's beliefs can be tested, shedding light on potential cognitive biases in their interaction, as studied in cognitive science and psychology [10]. The promise is that of a formally sound and empirically grounded approach that is better equipped than existing ones to deal with some central problems in social and formal epistemology.

Acknowledgements. We thank Lorenzo Casini, Davide Coraci, Daria Mikhaylova and the other members of the MInD research group at IMT

[13] We can conceive of the absence of a report that H in two ways: either the expert actually testified $\neg H$—*i.e.*, $\neg E_H$ will be equivalent to $E_{\neg H}$; or the expert did not assert anything about H at all.

Lucca for useful discussions. We acknowledge financial support from the MUR-PRO3 project "Understanding public data: experts, decisions, epistemic values" (CUP: D67G22000130001). Gustavo Cevolani acknowledges financial support from the Italian Ministry of University and Research (MUR) through the PRIN 2022 project "Reasoning with hypotheses: Integrating logical, probabilistic, and experimental perspectives" (grant n. 2022ARRY9N; CUP: D53D23009570006), funded by the European Union (Next Generation EU).

References

[1] C. Aitken, F. Taroni and S. Bozza, "Statistics and the Evaluation of Evidence for Forensic Scientists", Wiley, Hoboken, NJ, third edition edition, 2021.

[2] G. Betz, *Evaluating dialectical structures with Bayesian methods*, Synthese **163** (2008), 25–44.

[3] L. Bovens and S. Hartmann, "Bayesian Epistemology", Oxford University Press, Oxford, 2003.

[4] R. Brandom, "Making It Explicit: Reasoning, Representing, and Discursive Commitment", Harvard University Press, Cambridge, Mass., 1994.

[5] L. Casini and J. Landes, *Confirmation by robustness analysis: a Bayesian account*, Erkenntnis **89** (2024), 367–409.

[6] F. Claveau and O. Grenier, *The variety-of-evidence thesis: a Bayesian exploration of its surprising failures*, Synthese **196** (2019), 3001–3028.

[7] H. Collins and R. Evans, "Rethinking Expertise", University of Chicago Press, Chicago, 2019.

[8] V. Crupi, *Confirmation*, In: "The Stanford Encyclopedia of Philosophy", E. N. Zalta (ed.), Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

[9] C. Dahlman and L. Wahlberg, "Appeal to Expert Testimony – A Bayesian Approach", Springer, Cham, 2015, 3–18.

[10] I. E. Dror, *Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias*, Analytical Chemistry **92** (2020), 7998–8004.

[11] F. H. van Eemeren, "Argumentation Theory: A Pragma-Dialectical Perspective", Springer, Cham, 2018.

[12] A. Elga, *Reflection and disagreement*, Noûs **41** (2007), 478–502.

[13] R. Festa, *Testimonianze esperte e probabilità delle ipotesi*, L&PS – Logic & Philosophy of Science **XII** (2014), 3–39.

[14] R. Festa and G. Cevolani, *Unfolding the grammar of Bayesian confirmation: likelihood and antilikelihood principles*, Philos. Sci. **84** (2017), 56–81.

[15] B. Fitelson, *The plurality of Bayesian measures of confirmation and the problem of measure sensitivity*, Philos. Sci. **66** (1999), S362–S378.

[16] P. Garbolino, "Probabilità e Logica della Prova", Giuffrè Editore, Milano, 2014.

[17] G. Gigerenzer, "Calculated Risks: How To Know When Numbers Deceive You", Simon and Schuster, New York, 2002.

[18] A. I. Goldman, *Experts: Which ones should you trust?*, Philosophy and phenomenological research **63** (2001), 85–110.

[19] A. I. Goldman, "Why Social Epistemology is *Real* Epistemology", A. Millar A. Haddock and D. Pritchard (eds.), 2010, 1–29.

[20] I. Hacking, "An Introduction to Probability and Inductive Logic", Cambridge University Press, New York, 2001.

[21] U. Hahn, M. Oaksford and A. J. L. Harris, *Testimony and argument: a bayesian perspective*, In: "Bayesian Argumentation: the Practical Side of Probability", 2013, 15–38.

[22] C. L. Hamblin, "Fallacies", Vale Press, Newport News, Va., 1970.

[23] C. Howson and P. Urbach, "Scientific Reasoning: the Bayesian Approach", Open Court Publishing, Chicago, 2006.

[24] R. C. Jeffrey, "The Logic of Decision", University of Chicago Press, New York, 1965.

[25] R. C. Jeffrey, "Subjective Probability: The Real Thing", Cambridge University Press, Cambridge and New York, 2002.

[26] H. Jeffreys, "The Theory of Probability", Oxford University Press, Oxford, 1998.

[27] R. E. Kass and A. E. Raftery, *Bayes factors*, J. Amer. Statist. Assoc. **90** (1995), 773–795.

[28] P. J. Kellman, *et al.*, *Forensic comparison and matching of fingerprints: using quantitative image measures for estimating error rates through understanding and predicting difficulty*, PLOS ONE **9** (2014), 1–14.

[29] D. A. Lagnado, "Explaining the Evidence: How the Mind Investigates the World", Cambridge University Press, Cambridge, 2021.

[30] C. Martini, *Experts in science: a view from the trenches*, Synthese **191** (2014), 3–15.

[31] T. Mulligan, *The epistemology of disagreement: why not Bayesianism?*, **18** (2021), 587–602.

[32] B. Osimani and J. Landes, *Varieties of error and varieties of evidence in scientific inference*, British J. Philos. Sci. **74** (2023), 117–170.

[33] J. Sprenger and S. Hartmann, "Bayesian Philosophy of Science", Oxford University Press, Oxford and New York, 2019.

[34] J. M. Tangen, M. B. Thompson and D. J. McCarthy, *Identifying fingerprint expertise*, Psychological science **22** (2011), 995–997.

[35] B. T. Ulery *et al.*, *Accuracy and reliability of forensic latent fingerprint decisions*, Proc. Nat. Acad. Sci. India Sect. A **108** (2011), 7733–7738.
[36] D. Walton, “Appeal to Expert Opinion: Arguments from Authority”, Penn State Press, University Park, PA, USA, 2010.
[37] F. Zenker, ed., “Bayesian Argumentation”, Springer, Dordrecht, 2013.

Piero Avitabile
Gustavo Cevolani

Informational influence as a cause of (bi-)polarization? A simulative approach

1. *Introduction*

Lack of communication in a group may have dire consequences, especially when individuals adapt to others' behavior and ignore their private signals. Informational cascades [4] leading to the burst of market bubbles [11] are well-known dynamics where incomplete information may undermine individually rational decisions. Yet, the presence of barriers is not the only culprit of informational disasters. Even in situations of fully open communication, small errors may get magnified, as witnessed by many instances of *groupthink* studied in social sciences [18]. Informational drifts of the latter kind are often ascribable to dynamics of opinion polarization, which constitute our topic of interest.

Very often the opinions of entire groups radicalize in one direction after discussion. This dynamic is known as *group polarization* to social psychologists [17]. Polarization processes lie at the ground of many divides in public debate, where the opinion of individuals about a given issue moves towards opposite poles of the opinion spectrum, thus forming two distant cohesive clusters. It is in fact this bimodal distribution of opinions which is commonly referred to as "polarization" in everyday discourse [3]. For the sake of clarity, we instead refer to the latter as *bi-polarization*, to indicate that this is the effect of two diverging polarization dynamics.

The present work builds upon [28, 29], extending previous one by [23], and further explores the conditions under which exchange of arguments among agents can generate bi-polarization effects. Our main tool are Agent Based Models (ABMs) of opinion dynamics. ABMs are computer models that simulate a society of agents interacting with their neighbors through some communication link. At an abstract level, the group is represented as a network where the nodes are agents and the edges are ways by which influence unfolds. Agents are artificial entities: their behavior and modes of interaction are programmable, possibly with given amounts of randomness and different initial parameters. In the more restricted case of ABMs of opinion dynamics, interaction among agents consists of *information exchange* inducing an opinion update for the receiver. In our specific type of models, information consists of *arguments* for or against a debated issue (see Section 2 for more methodological details).

Previous work in [29] shows two main things, namely:

(a) that the relative strength of the arguments employed in a debate may have a strong effect on bi-polarization dynamics, driving more individuals to end up on the stronger side of the opinion spectrum;
(b) that informational biases can induce bi-polarization, although under strong conditions.

Here, by "informational bias" we indicate the general tendency of individuals to prioritize information in one direction rather than the other (typically favoring information that resonates with one's prior opinion). In our setting, this may happen either by agents communicating a specific kind of arguments, or else by updating their opinion in a specific way upon receiving arguments from someone else (see Section 3.2).

In the present work we revise the assumptions and experiments conducted to show (a) and (b), in order to test their robustness against relevant additional factors that may alter the dynamics of information exchange. First, we show that results concerning (a) are strongly influenced by the specific measure of argument strength that is adopted. As a second point, we also show that the presence of external influence (*e.g.* "propaganda") on both sides facilitates bi-polarization effects, therefore mitigating the conditions for their occurrence postulated in (b). As a third aspect, we also investigate whether and if some initial degree of bi-polarization may have beneficial truth-tracking effects for the group as a whole. In philosophy of science an analogous hypothesis is known as the "benefit of transient diversity" after [36] and is used to explain the added value of diversity on scientific discovery. Initial results in this direction show that this is in fact the case, and the reasons are interesting to expose.

This paper proceeds as follows. In Section 2 we provide an overview and state of the art on agent-based modelling of opinion dynamics, with special focus on models based on argument exchange. Section 3 introduces the model by [23] and its extensions by [29]. Section 4 illustrates the new experiments and their results.

2. *Related work*

Multi-agent models of opinion dynamics have been employed to understand how mutual influence among individuals in a communication network may lead to opinion consensus or to fragmentation into subgroups with alternative views [1, 9, 12, 13, 15, 16]. ABMs are a useful complement to empirical research in situations that lab and field experiments cannot cope with. This holds in particular when handling large groups and a large number of parameters. In fact, multiple simulation runs can provide statistically significant results in a controlled setting, therefore avoiding many of the replicability issues that determine the so-called *reproduction crisis* in psychology and social sciences [32].

In the first-generation of models of opinion dynamics, the opinion of one agent at time $t+1$ is obtained by averaging his prior opinion at t with those of his communication partners. As such, these models can hardly account for bi-polarization effects. In fact, the mechanism of averaging cannot reproduce by itself the radicalization dynamic occurring with group polarization [23].

As a further important aspect, agents of first-generation ABMs only display their opinion to others, not their motivations for it. Such mechanism of influence is too coarse-grained to reproduce one informational trigger which has been proven, by experimental studies in social psychology, effective for the emergence of polarization, namely the exchange of novel and persuasive *arguments* in one direction or the other [22, 25, 34, 35]. More recent models integrate *argument communication* as a main element of influence [14, 23], assuming that the opinion of one agent on a debated issue is determined by the arguments that she considers relevant and, furthermore, that agents communicate such arguments during interaction. First results with these models show that social influence of this kind can generate bi-polarization effects by just postulating a moderate degree of homophily, that is the tendency to communicate with individuals having similar opinions [6, 23, 24]. One limitation of these models is that they assume that different arguments are unrelated, i.e they do not support, undermine or rebut each other. This limitation does not allow to account for how one argument can strengthen or weaken another, and, in turn, to provide a measure of *argument strength*, as relevant in this context, unless arbitrarily as in [31].

A new family of models [2, 29, 30, 33] adopts the more structured approach of abstract argumentation [10] to account for this dimension. In particular, the present work capitalizes on the richer formalism of argument graphs with both *attacks* and *supports* between arguments [8]. Arguments are further endowed with a measure of strength, as defined in *gradual* argumentation [7]. To make the use of argument graphs more transparent, let us consider the following example of a fictional debate between Alice and Bob.

Example 2.1. Alice: We should buy a house with a garden near the park, because it is a very nice area. (a)

Bob: There are swamps with a lot of mosquitoes in the park, therefore it is not such a nice area. (b)

Alice: the area is regularly disinfested and therefore mosquitoes are not a threat. (c)

Bob: They are going to build a playground for kids, therefore the area will become even nicer and apt for families. (d)

Alice: The decision of building a playground is still in the making, therefore it is not sure that there will be one. (e)

Arguments from a to e are represented as nodes in the graph of Figure 1. Clearly, argument b attacks a by undermining its premise. This is represented in the graph by a single arrow from b to a. In turn, argument c attacks b and therefore *defends* a from b. Intuitively, c has a positive influence on a since it reinstates it. Argument d instead provides support for a, somehow reinforcing its premise. Here, support is represented graphically by a double-edged arrow from d to a. Again, d is attacked by e, which therefore also affects a in a negative way. Our measures of argument strength presented in Section 3.1 will take into account all these positive and negative influences determined, both directly and indirectly, by attacks and supports.

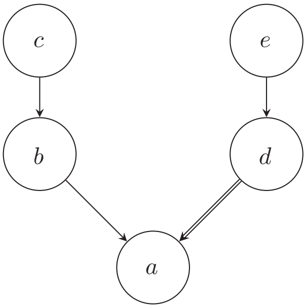


Figure 1 An example of argument graph with attacks and support for Example 2.1. Labelled nodes represent arguments. Relations of attack between arguments are indicated with a single arrow and relations of support with a double arrow.

3. *Model description*

The ABM employed here extends the one presented in [29], which in turn builds upon the model of *Argument Communication Theory of Bipolarization* (ACTB) by [23]. Here, we describe its general architecture and additions for the present work. For more details on definitions and implementation we refer to [23] and [29]. The code is available at https://www.comses.net/codebase-release/ea5f9a4e-321a-4c39-a5e2-18ba7657ed82/.

3.1. *Basic components*

Our ABM consists of a society of n agents and a finite directed, connected and acyclic graph $\mathbf{G} = \langle G, R^-_\mathbf{G}, R^+_\mathbf{G}\rangle$ representing the *global knowledge base* for the debated issue v. G is the set of all arguments about v, the latter being included as the terminal node of the graph. The relation $aR^-_\mathbf{G}b$ (resp. $aR^+_\mathbf{G}b$) indicates an attack (resp. support) from argument a to argument b. Intuitively, the global knowledge base G represents the whole information potentially available about v, condensating arguments (in favor and against v) and their mutual relationships.

At each point t in time, every agent i owns a fixed number of arguments (the same number for all agents), which is a subset $S_{i,t}$ of **G**, ordered by recentness. Such arguments inform one agent's *individual knowledge base* $\mathbf{S_{i,t}}$ as a (proper) induced subgraph of G.[1] Arguments owned by the agents determine their opinion as specified in the subsequent steps.

Arguments in a graph influence each other, directly and indirectly, via attacks and supports, and such influence may be either negative or positive. We denote by $Neg_\mathbf{A}(a)$ the set of arguments with a negative influence on argument a relative to graph **A**. This set is constituted by all arguments linked to a via a directed path containing an *odd* number of R^- transitions. In the graph of Figure 1 these arguments are b and e. On the contrary, the set $Pos_\mathbf{A}(a)$ is the set of arguments with a positive influence on a and is constituted by all arguments linked to it by directed path with an *even* number (including 0) of R^- transitions. This is the case of c and d in Figure 1

Based on this distinction, in [29] we provide a measure of *argument strength* for each node, which in turn serves to determine each agent's opinion. To this end, we endow our graphs with a function $w_\mathbf{A} : A \longrightarrow [0, 1]$ assigning to each argument a base score, which is always 1 in our simulations. The base score provides the first step for iteratively computing the proper strength of any argument based on the strengths of its ancestors (if any). The strength $s_{\mathbf{S_{i,t}}}(a)$, of argument a relative to the individual

[1] To be precise, the individual knowledge base of i at t is $\mathbf{S_{i,t}} = \langle S_{i,t}, R^-_{\mathbf{S_{i,t}}}, R^+_{\mathbf{S_{i,t}}}\rangle$, where $S_{i,t} \subseteq G$ always contains the topic node v; and where $R^-_{\mathbf{S_{i,t}}} = R^-_\mathbf{G} \cap (S_{i,t} \times S_{i,t})$ and $R^+_{\mathbf{S_{i,t}}} = R^+_\mathbf{G} \cap (S_{i,t} \times S_{i,t})$ are the restrictions of the support and attack relations of G to the arguments of $S_{i,t}$.

knowledge base of agent i at time t, is defined as[2]:

$$s_{\mathbf{S_{i,t}}}(a) = \begin{cases} w_{\mathbf{S_{i,t}}}(a) & \text{if } \star \\ \frac{1+\frac{\sum_{b \in S_{i,t} \cap Pos_{\mathbf{G}}(a)} s_{\mathbf{S_{i,t}}}(b) - \sum_{b \in S_{i,t} \cap Neg_{\mathbf{G}}(a)} s_{\mathbf{S_{i,t}}}(b)}{|S_{i,t} \cap Pos_{\mathbf{G}}(a)| + |S_{i,t} \cap Neg_{\mathbf{G}}(a)|}}{2} & \text{otherwise} \end{cases} \tag{3.1}$$

where $\star$ stands for $(Pos_{\mathbf{G}}(a) \cup Neg_{\mathbf{G}}(a)) \cap S_{i,t} = \emptyset$. Therefore, the strength of leaf nodes (*i.e.* arguments with no ancestors) is calculated as their base score, while the strength of other arguments is provided by weighting their ancestors with positive influence against those with negative influence. This measure ranges from 0 (weakest) to 1 (strongest). The agent's i opinion $o_{i,t}$ at time t is then defined simply as the strength of the topic node, *i.e.*

$$o_{i,t} = s_{\mathbf{S_{i,t}}}(v) \tag{3.2}$$

Here, $o_{i,t} = 0$ means that the opinion of i is totally negative, while $o_{i,t} = 1$ means totally positive.

3.2. *Interaction*

The agents' opinions evolve as the result of exchange of arguments among randomly selected communication partners in a sequence of interactions. At every time step two things occur:

Coupling. Agent i (the receiver) is randomly selected and paired with a communication partner j (the sender). The latter is selected with a probability proportional to the similarity of its opinion with that of i at time t, noted as $sim_{i,j,t}$ - ranging from 0 (mostly dissimilar) to 1 (mostly similar). This probability is calculated as:

$$p_{j,t} = \frac{(sim_{i,j,t})^h}{\sum_{j=1, j \neq i}^{n} (sim_{i,j,t})^h} \tag{3.3}$$

Here, h implements homophilous selection. With $h = 0$ the receiver is equally likely to be paired with anyone else as a sender. As the value of h increases, the agent is more and more likely to be coupled with someone having similar opinions.

[2] Note that $s_{\mathbf{S_{i,t}}}$ is well-defined only for acyclic graphs, but this is enough for our purposes and the simulations we conduct.

Informational influence. The sender j communicates some argument from its individual knowledge base — *i.e.* one from $S_{j,t}$ — and the receiver i processes it and then revises its opinion after modifying its individual knowledge base. In the original ACTB model the *standard* protocol of informational influence at t contemplates that the sender j communicates one random argument, and the receiver i updates her knowledge base to $S_{i,t+1}$ by adding this argument as the most recent, while forgetting the oldest in $S_{i,t}$. In the standard protocol there is no filter on which arguments are communicated and which ones are accepted. In [29] we add alternative procedures to the standard protocol, implementing biased communication (by the sender) and biased update (by the receiver), in order to test their bi-polarizing effect.

On the sender's side, we implemented two different forms of biased communication, namely:

1. *Preferential communication of arguments favouring one's prior opinion (PCO).* The sender communicates some (randomly selected) argument that is in line with his current opinion, *i.e.* a positive argument if its opinion is favorable (above 0.5), and a negative one otherwise.

and

2. *Preferential communication of stronger arguments (PCS).* The sender j communicates some random argument among those that he considers stronger according to the function $s_{\mathbf{S_{j,t}}}$.

PCO encodes a rather partisan behaviour, for it assumes that the sender only shares information "that he likes". On the other hand, PCS can be regarded as a more "honest" attitude, since the agent communicates the pieces of information he considers to be, to the best of her knowledge, the most justified.

On the receiver's side, we instead tested additional modalities of update, the first one is

3. *Preferential update with arguments favouring one's prior opinion (PUO).* The receiver rejects any argument not in line with his current opinion, *i.e.* rejects a positive argument if his opinion is negative (below 0.5), and a negative one if positive. Otherwise he follows the standard procedure.

This updating modality is the analogous of PCO, whereby the agent simply obliterates information she does not like. This is probably the most immediate way of implementing the psychological mechanism of *confirmation bias* [26].

A different updating modality is

4. *Preferential discarding of weaker arguments (PUW)*. The receiver i discards the oldest argument among those that he considers as weaker according to the function $s_{\mathbf{S_{i,t}}}$.

As for PCS, here the agent arguably acts "fairly" when obliterating arguments that she is entitled to consider as weak.[3]

Finally, we implemented

5. *Vigilant update with n new arguments (VUn)*. The receiver accepts any argument against his opinion, but also adds n new arguments favoring his opinion.[4] He instead follows the standard protocol when receiving an argument supporting his opinion.

This procedure describes an alternative implementation of the mechanism of confirmation bias. In fact, it encodes instead a more proactive attitude, also well-known to psychologists, to contrast dissonant information by searching for new confirming pieces of evidence or justificatory reasons [27]. The parameter n indicates how strong such contrast is.

In all these updating procedures the size of the individual knowledge base is kept constant (and limited), as assumed in the original model of [23], on the basis of psychological research on memory processes [5]. For *VUn* this is done by forgetting the $n + 1$ oldest arguments instead of just one.

For all policies from 1. to 5. it is possible to vary the probability P by which they are implemented at each time step, with respect to the alternative of following the standard protocol of communication or update. For example, $P(PCO) = x$, with $0 \leq x \leq 1$, means that the sender j uses PCO with probability x, and the standard communication protocol with probability $1 - x$. Therefore, P measures the strength of the bias.

In the original ACTB model there are two equilibria of the system, namely *perfect consensus* and *maximal bi-polarization*. Such equilibria occur when a sufficient number of arguments gets forgotten by all members of the group and therefore cannot circulate anymore. Perfect consensus means

[3] We let the agent forget the oldest argument among the weaker ones in order to stick as much as possible to the updating protocol of the original model, and for no other reason. In fact, there is room for arguing that discarding an argument which survived for a longer period may after all not be "rational".

[4] In [29] new arguments are selected by the agent among all potential argument (the whole knowledge base G), thus simulating a process of individual inquiry. An alternative option would be to let the agent draw them out from the more restricted set of arguments that are still alive in the community. Intuitively, this option amounts to "asking someone else" for confirmation.

all agents end up having the same arguments and therefore the same opinion. Here, polarization scenarios where all agents end up with the same fully positive (resp. fully negative) opinion, count as a specific form of consensus.

Maximal bi-polarization instead means the emergence of two subgroups, one where agents have opinion 1 owning the same positive arguments, and one where agents have opinion 0 owning the same negative ones. In the original model equilibria are stable, in the sense that no further change in individual knowledge bases is possible, and therefore no individual opinion change.[5]

3.3. *Additions*

Modifying the measure of argument strength. As mentioned in the introduction, one point of this work is to test the robustness of previous results under modifications of the measure of argument strength. There are many reasonable alternatives to the measure provided in Equation 3.1, and which have intuitive grounding. Here we implement one of them. To understand its definition and rationale, it is key to note that in Equation 3.1 ancestors with influence on a are relative to the global knowledge base, *i.e.* they belong to either $Pos_{\mathbf{G}}(a)$ or $Neg_{\mathbf{G}}(a)$, and not relative to the individual knowledge base of the agent, *i.e.* $Pos_{\mathbf{S_{i,t}}}(a)$ or $Neg_{\mathbf{S_{i,t}}}(a)$. In fact, the graph $\mathbf{S_{i,t}}$ may be easily disconnected, while $\mathbf{G}$ is not. Therefore the measure of Equation 3.1 allows to count arguments with no path to v in the agent's individual knowledge base. It then implicitly assumes that the agent knows that, say, some argument c defending a against a counterargument b has positive influence towards a although the agent is not aware of b. This assumption may be very strong in certain contexts of application, especially those where there is no intuitive link between c and a. We can revise this assumption by replacing, in Equation 3.1, the following:

$$\frac{1+\frac{\sum_{b\in Pos_{\mathbf{S_{i,t}}}(a)} s_{\mathbf{S_{i,t}}}(b)-\sum_{b\in Neg_{\mathbf{S_{i,t}}}(a)} s_{\mathbf{S_{i,t}}}(b)}{\max\{1,|Pos^{+}_{\mathbf{S_{i,t}}}(a)|+|Neg^{+}_{\mathbf{S_{i,t}}}(a)|\}}}{2} \tag{3.4}$$

where $Pos^{+}_{\mathbf{S_{i,t}}}$ (resp. $Neg^{+}_{\mathbf{S_{i,t}}}$) is the set of arguments with positive (resp. negative) influence w.r.t. the graph $\mathbf{S_{i,t}}$ and whose strength is > 0. Overall,

[5] It is important to note that bi-polarization is an equilibrium only when $h \neq 0$. Otherwise, communication and opinion change are still possible between maximally distinct subgroups. Further, our specific implementation of vigilant update (VUn) allows that forgotten arguments can possibly circulate again, therefore undermining the stability of the system. Many of our experiments are run under the condition $h = 0$ or use VUn. In this cases we check that the halt conditions we impose are sufficient to guarantee stability in the limit.

this equation replaces all occurrences of $Pos_{\mathbf{G}}$ (resp. $Neg_{\mathbf{G}}$) with $Pos_{\mathbf{S_{i,t}}}$ (resp. $Neg_{\mathbf{S_{i,t}}}$) in the numerator, and with $Pos^{+}_{\mathbf{S_{i,t}}}$ (resp. $Neg^{+}_{\mathbf{S_{i,t}}}$) in the denominator. As a consequence, all disconnected arguments and those with null strength get discounted.[6] Then we get a measure s^{*} alternative to that of Equation 3.1. As we shall see in Section 4.1 this simple modification has a strong impact on our previous results.

External influence. Thus far, all versions of our model assume that agents exchange information (in a more or less biased way) only with their peers, *i.e.* other individuals subject to influence and opinion change. This overlooks a relevant aspect of daily life social contexts, which is *influence* from *external* actors. The latter may take many forms, ranging from influence in one direction and by a single source of information (*e.g.* propaganda) to multiple voices pointing to alternative directions (*e.g.* trendsetters or influencers). What is common to all these sources is that they are assumed not to be subject, themselves, to influence and opinion change. Abstracting away from other differences, in our model we encode the effect of (positive or negative) propaganda simply as a probability P' that the receiver, once selected, is not paired with another agent but is instead exposed to listening arguments in a specific direction.

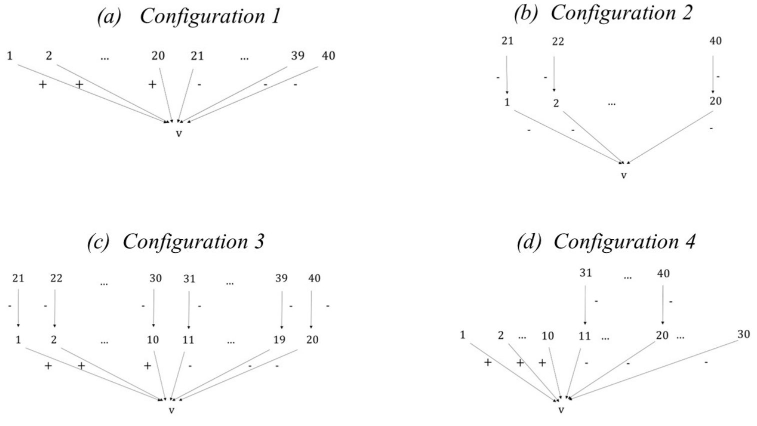
Varying homophily during a single run. A further characteristic of the architecture employed thus far is that homophily is kept constant during the whole run. This sounds as a reasonable assumption if individuals are supposed to be, at any moment, totally free to interact with others as they prefer. For many purposes, it seems however reasonable to test the effect of an external intervention in this regard. As an example, it is claimed from many sides that debate and decision-making on sensible topics may benefit by fostering diversity between opinions and listening to alternative views. This foreshadows the opportunity of an artificial intervention, during specific phases, redirecting the agents' propensity to interact with whoever they like. Here we act on the level of the agents' homophily in order to test the benefit of transient diversity (Section 1). The specific intervention here consists of setting an initial high level of homophily and then removing it, as we specify further below.

[6] The value $\max\{1, |Pos^{+}_{\mathbf{S_{i,t}}}(a)| + |Neg^{+}_{\mathbf{S_{i,t}}}(a)|\}$ in the denominator serves to avoid indeterminate forms when all ancestors are disconnected or have null strength. In these cases, it is assumed that the agent's opinion should be 0.5.

4. *Simulations and results*

4.1. *Discounting disconnected arguments*

A first question investigated in [29] was whether a generally positive (resp. negative) attitude towards the topic v may be induced by stronger arguments for (resp. against) v. More precisely, we tested whether making arguments from $Pos_{\mathbf{G}}(v)$ stronger, while equal in number, than arguments in $Neg_{\mathbf{G}}(v)$, would induce a higher rate of consensus on a positive opinion or else, when bi-polarization occurs, force more agents to polarize towards the positive side, i.e having opinion 1. To test this, in [29] we ran our simulation with a society of 20 agents. We compared two different configurations of the global knowledge base, namely Configuration 1 and Configuration 2 of Figure 2. Both configurations have an equal number of positive and negative arguments ($Pos_{\mathbf{G}}(v) = Neg_{\mathbf{G}}(v) = 20$). In Configuration 1, positive and negative arguments are also equally strong, since all of them are unattacked (and have the same base score of 1). This is reflected by the fact that the strength $s_{\mathbf{G}}(v)$ of the topic node here is 0.5, *i.e.* right in the middle of the opinion spectrum. In Configuration 2 the positive arguments (unattacked) are instead stronger than the negative ones, and in fact $s_{\mathbf{G}}(v)$ is 0.75.


(a) Configuration 1

(b) Configuration 2

(c) Configuration 3

(d) Configuration 4

FIGURE 2 Four different configurations of the global knowledge base, all with $Pos(v) = Neg(v) = 20$. The strength of the topic node for each of these global knowledge bases varies from $s_{\mathbf{G}}(v) = 0.5$ for (a) and (c), to $s_{\mathbf{G}}(v) = 0.625$ for (d) and $= 0.75$ for (b).

In [29] we conducted our tests with a high level of homophily ($h = 9$) and by varying the memory of agents from $|S_{i,t}| = 4$ to $|S_{i,t}| = 8$, where the former means that agents have only four relevant arguments (besides v) in their individual knowledge base, while they have eight in the latter. As expected, under the first configuration the average opinion after multiple runs was constantly around 0.5 and, in cases of bi-polarization, the average size of pro-oriented and con-oriented groups was the same (*i.e.* 10). Things change with Configuration 2: although we did not witness significant opinion shifts in cases of consensus, for bi-polarizations the average size of the pro-oriented group increased significantly above the mean, with a peak of 11.56 for $|S_{i,t}| = 8$ (high individual memory). Overall, this seems to witness that stronger arguments for one side can tilt the balance towards that direction.

However, further experiments conducted for the present work put this conclusion in a more detailed perspective. As a test, we modified our measure of argument strength with the new s^* as indicated in Equation 3.4. Here, we run our tests with Configuration 4 of Figure 2 as the global knowledge base, where $s^*_{\mathbf{G}}(v) = 0.66$ and $s_{\mathbf{G}}(v) = 0.625$, *i.e.* it is still positive both with the new and the old measure.[7] We assume $|S_{i,t}| = 8$ and $h = 9$. Results are reported in Figure 3. With the old measure $s_{\mathbf{S_{i,t}}}$, the outcome is very similar to the one obtained with Configuration 2, where the pro-oriented group dominates. On the contrary, with the new measure $s^*_{\mathbf{S_{i,t}}}$ the situation is inverted, with the average size of the pro-oriented group falling way below the medium. This was expected, since, given that stronger arguments gets easily discounted in the individual opinion, due to their being disconnected in the individual knowledge base, the dynamic is specular to what happens with the old measure (see [29, par. 4.6]).

4.2. *External influence as propaganda*

A second target in [29] was to ascertain whether bi-polarization effects could be generated without homophily ($h = 0$) – or with a low level of it ($h = 1$) – as the effect of biases in communication and update. To test this, we worked mainly with Configuration 1 as the global knowledge base, and results where rather surprising. Except for situations where agents systematically reject contrary arguments ($P(PUO)$ = 1), bi-polarization effects hardly obtain. In fact, having tested various combinations of biased communication (1. and 2.) and update (3.-5.) with different levels of probability, the most typical pattern that emerged was simple polarization, with a dynamic similar to that represented in Figure 4. Here, after an initial phase of splitting into two groups, one group gets progressively attracted

[7] We run this experiment on a different global knowledge base in order to avoid a natural problem that our new measure generate w.r.t. Configuration 2, namely that no agent can get a totally positive opinion, because all positive arguments would be disconnected from the topic node in their individual knowledge base.

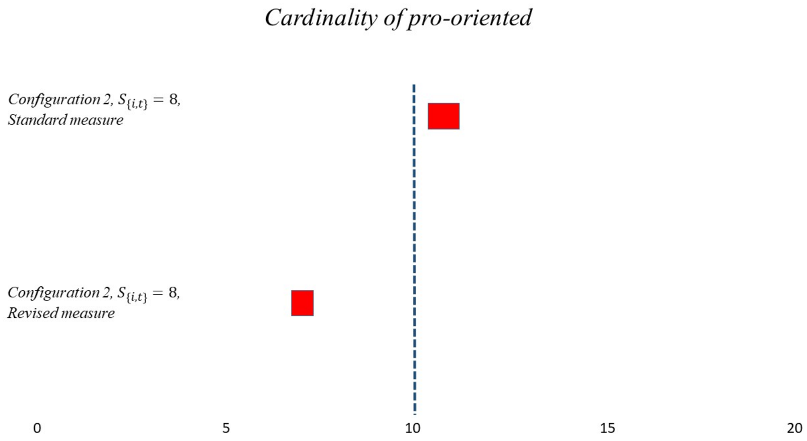


FIGURE 3 Average size of the pro-oriented group after bi-polarization in Configuration 4 of Figure 2, with the old measure $s_{\mathbf{S}_{i,t}}$ (above) and the new one $s^*_{\mathbf{S}_{i,t}}$ (below). With $h = 9$ and the old measure the average cardinality is ~ 10.78 with a 95% confidence interval of width 0.42 (*i.e.* ranging from 10.35 to 11.2). With the new measure it is ~ 7.35, with a 95% confidence interval of width 0.35.

towards the opposite pole. This happens even when agents have a consistent probability of communicating arguments in favour of their opinion ($P(PCO) = 0.9$) and of rejecting contrary arguments ($P(PUO) = 0.9$).

To test the robustness of this dynamic, we added external influence in the form of of propaganda from both sides of the opinion spectrum, and things change quite substantially. In this specific setting, we add external influence (as generically described in Section 3.3) in the form of agents facing communication from an external source which only voices arguments that support their current opinion. We tested this against a previous batch of simulations where bi-polarization occurred only in 12% of the runs with $P(PUO) = P(PCO) = 0.9$ – with no homophily and low individual memory ($|S_{i,t}| = 4$).

Our new simulations show that with a probability of hearing favorable propaganda of only 0.05% the number of runs ending in bi-polarization goes up to 23.6% (118 over 500 runs). With a probability of 0.15, they raise to more than 60% (306 over 500), and to almost 97% (484 over 500) with a probability of 33%.

To check the real impact of this type of external influence, we repeated the same simulations without biased agents, *i.e.* under the standard protocol with $P(PUO) = P(PCO) = 0$. With 0.05% probability of external influence, the result is quite similar to the biased case with 106 over

500 runs ending in bi-polarization. However, raising the probability of external influence to 0.15% and 0.33% makes bi-polarizations disappear, forcing again polarization dynamics akin to that in Figure 4.

These results show that even a small amount of external influence can disrupt the polarization dynamic described above and facilitate bi-polarization instead. In fact, low levels of it cause bi-polarizing effects also without bias. This however does not exclude bias as a necessary cause of bi-polarization in cases where external influence is higher. It follows that, by assuming propaganda as a relevant factor in our model, confirmation bias can be seen as a concurrent cause of bi-polarization effects, when acting in tandem with external influence. To some extent, these results restore the widespread intuition that confirmation bias is a responsible for this phenomenon [21].

4.3. *The benefit of transient diversity*

Many historical examples witness that diversification of efforts and a certain degree of pluralism has a beneficial effect on scientific discovery [19, 20]. Although results need to converge and be compared at some point, scientific inquiry may cash in on phases where independent lines of research run in parallel. Work in [36] provides a formal model to show the

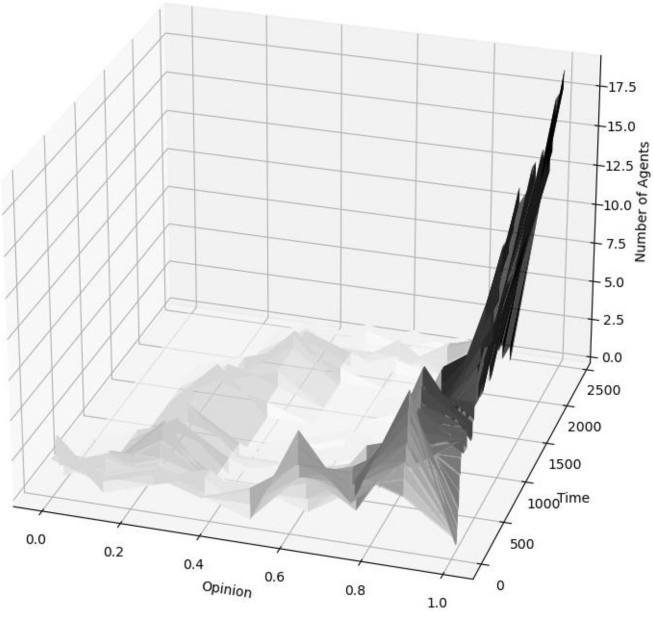


FIGURE 4 Typical dynamic of polarization with bias, no homophily and no external influence.

benefit of *transient diversity* in a community, the latter being understood as a situation where "diversity should be around long enough so that individuals do not discard theories too quickly, but also not stay around so long as to hinder the convergence to one action."

In a model like ours, it becomes interesting to test whether fostering a temporary divergence of opinions is somehow truth-conducive. One first problem with our model is that *truth*, as commonly understood, is not defined: all there is are arguments and opinions. Yet, the measure $s_{\mathbf{G}}(v)$, *i.e.* the strength of the topic node according to the whole global knowledge base, provides something of this kind, that is the value of the opinion that an ideal agent should hold on the basis of the most complete information available. For our experiment we chose Configuration 2 of Figure 2, where $s_{\mathbf{G}}(v) = 0.75$.

For a first test on the effect of transient diversity in our framework, we run our simulations under the standard communication protocol and high individual memory of $S_{i,t} = 8$, by setting a high level of homophily ($h = 9$) until bi-polarization occurs, and then moving it away ($h = 0$) to force consensus. We then calculate the average opinion in all runs ending with consensus. Finally, we compare these results with those from analogous setups but with constant level of homophily, either constantly high ($h = 9$) or absent ($h = 0$).

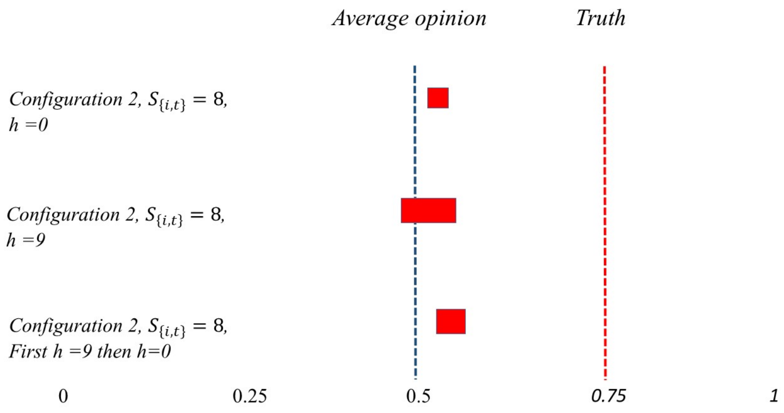


FIGURE 5 Average opinion after consensus for Configuration 2 ($S_{i,t} = 8$) with constant homophily ($h = 0$, above) and ($h = 9$, middle) and with transient diversity (first $h = 9$ and then $h = 0$, bottom). With $h = 0$ the average opinion is ~ 0.53 with a 95% confidence interval of width 0.01 (*i.e.* ranging from 0.52 to 0.54). With $h = 9$ the average opinion is ~ 0.52, but with a 95% confidence interval of width 0.03. With transient diversity the average opinion is ~ 0.55 with a 95% confidence interval of width 0.025.

As illustrated in Figure 5, our results over 500 simulations show that the average consensus opinion in situations of transient diversity comes slightly closer to the actual value of $s_{\mathbf{G}}(v)$, moving away from the median opinion in a way that is statistically significant. This indicates that, in this model, transient diversity has, at least for some configurations, a veritistic value. This result is indeed coherent with our previous results from [29] on the effect of argument strength (see Section 4.1). In fact, since a larger number of agents polarizes towards a totally positive opinion than towards a negative one, it is to be expected that the larger group works as an attractor after barriers are removed.

5. *Conclusions*

This paper extends previous work of [28, 29] by testing a new measure of opinion and by encoding new features such as the effect of propaganda and of transient diversity.

Changing the measure has the significant effect of reversing the balance of power between groups in bi-polarization. With the new measure, the group with weaker reasons ends up getting larger. Neither this nor the previously adopted measure are empirically grounded, and therefore none of them speaks in favor of a specific trend. Yet, the second way of measuring argument strength seems reasonable in contexts where indirect dependencies between arguments are complex to assess. Therefore, our results provide a specific warning about quickly optimistic conclusions on the truth conducive power of argument strength.

One second batch of simulations tested the bi-polarizing effect of propaganda in biased communication. Our results witness that even a small degree of external influence can force a significant level of bi-polarization, although different combinations of their levels may generate opposite outcomes.

Finally, we tested whether an initial phase of high homophily, with consequent bi-polarization, can make the group's opinion more accurate when forced to reconverge. Although still partial, our results seem to indicate a slight benefit with respect to situations of fully free circulation of information (no homophily), and a more significant one w.r.t. situations where homophily canalizes social interaction.

Acknowledgements. The infrastructure for the simulative part of this work was provided by the GARR Consortium. Research by Davide Chiarella was funded by the European Union – NextGenerationEU and by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.5, project “RAISE – Robotics and AI for Socio-economic Empowerment” (ECS00000035).

References

[1] R. P. Abelson, *Mathematical models in social psychology*, In: “Advances in Experimental Social Psychology”, Vol. 3, Elsevier, 1967, 1–54.

[2] G. Betz, *Natural-language multi-agent simulations of argumentative opinion dynamics*, arXiv preprint arXiv:2104.06737, 2021.

[3] A. Bramson, P. Grim, D. J. Singer, W. J. Berger, G. Sack, S. Fisher, C. Flocken and B. Holman, *Understanding polarization: Meanings, measures, and model evaluation*, Philos. Sci. **84** (2017), 115–159.

[4] S. Bikhchandani, D. Hirshleifer and I. Welch, *A theory of fads, fashion, custom, and cultural change as informational cascades*, Journal of political Economy **100** (1992), 992–1026.

[5] G. D. A. Brown, I. Neath and N. Chater, *A temporal ratio model of memory*, Psychological review **114** (2007), 539.

[6] S. Banisch and E. Olbrich, *An argument communication model of polarization and ideological alignment*, arXiv preprint arXiv:1809.06134, 2018.

[7] P. Baroni, A. Rago and F. Toni, *From fine-grained properties to broad principles for gradual argumentation: A principled spectrum*, Internat. J. Approx. Reason. **105** (2019), 252–286.

[8] C. Cayrol and M.-C. Lagasquie-Schiex, *Bipolar abstract argumentation systems*, In: “Argumentation in Artificial Intelligence”, Springer, 2009, 65–84.

[9] G. Deffuant, D. Neau, F. Amblard and G. Weisbuch, *Mixing beliefs among interacting agents*, Adv. Complex Syst. **3** (2000), 87–98.

[10] P. Minh Dung, *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*, Artificial intelligence **77** (1995), 321–357.

[11] D. Easley and J. Kleinberg, “Networks, Crowds, and Markets: Reasoning About a Highly Connected World”, Cambridge University press, 2010.

[12] J. R. P. French Jr., *A formal theory of social power*, Psychological review **63** (1956), 181.

[13] N. E. Friedkin and E. C. Johnsen, *Social influence and opinions*, J. Math. Sociol. **15** (1990), 193–206.

[14] N. E. Friedkin, A. V. Proskurnikov, R. Tempo and S. E. Parsegov, *Network science on belief system dynamics under logic constraints*, Science **354** (2016), 321–326.

[15] F. Harary, *A criterion for unanimity in french’s theory of social power*, (1959).

[16] R. Hegselmann, U. Krause *et al.*, *Opinion dynamics and bounded confidence models, analysis, and simulation*, Journal of artificial societies and social simulation **5** (2002).

[17] D. J. Isenberg, *Group polarization: A critical review and meta-analysis*, Journal of personality and social psychology **50** (1986), 1141.

[18] I. L. Janis, "Groupthink", Houghton Mifflin Boston, 1983.

[19] T. S. Kuhn, "The Structure of Scientific Revolutions", University of Chicago press, 2012.

[20] I. Lakatos and P. Feyerabend, "For and Against Method: Including Lakatos's Lectures on Scientific Method and the Lakatos-Feyerabend Correspondence", University of Chicago Press, 2019.

[21] C. G. Lord, L. Ross and M. R. Lepper, *Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence*, Journal of personality and social psychology **37** (1979), 2098.

[22] D. G. Myers and G. D. Bishop, *Discussion effects on racial attitudes*, Science **169** (1970), 778–779.

[23] M. Mäs and A. Flache, *Differentiation without distancing. explaining bi-polarization of opinions without negative influence*, PloS one **8** (2013), e74516.

[24] M. Mäs, A. Flache, K. Takács and K. A. Jehn, *In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization*, Organization science **24** (2013), 716–736.

[25] D. G. Myers, *Polarizing effects of social interaction*, Group decision making **125** (1982), 137–138.

[26] R. S. Nickerson, *Confirmation bias: a ubiquitous phenomenon in many guises*, Review of general psychology **2** (1998), 175–220.

[27] R. E. Nisbett and T. D. Wilson, *Telling more than we can know: Verbal reports on mental processes*, Psychological review **84** (1977), 231.

[28] C. Proietti and D. Chiarella, *Measuring bi-polarization with argument graphs*, In: "AI^3@ AI* IA", 2021.

[29] C. Proietti and D. Chiarella, *The role of argument strength and informational biases in polarization and bi-polarization effects*, Journal of Artificial Societies and Social Simulation **26** (2023), 5.
URL: http://jasss.soc.surrey.ac.uk/26/2/5.html,
https://doi.org/10.18564/jasss.5062 doi:10.18564/jasss.5062.

[30] C. Proietti, *The dynamics of group polarization*, In: "Logic, Rationality, and Interaction: 6th International Workshop, LORI 2017, Sapporo, Japan, September 11-14, 2017, Proceedings 6", Springer, 2017, 195–208.

[31] D. J. Singer, A. Bramson, P. Grim, B. Holman, J. Jung, K. Kovaka,

A. Ranginani and W. J. Berger, *Rational social and political polarization*, Philos. Stud. **176** (2019), 2243–2267.

[32] J. P. Simmons, L. D. Nelson and U. Simonsohn, *False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant*, Psychological science **22** (2011), 1359–1366.

[33] P. Taillandier, N. Salliou and R. Thomopoulos, *Introducing the argumentation framework within agent-based models to better simulate agents' cognition in opinion dynamics: Application to vegetarian diet diffusion*, Journal of Artificial Societies and Social Simulation **24** (2021).

[34] A. Vinokur and E. Burstein, *Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach*, Journal of Personality and Social Psychology **29** (1974), 305.

[35] A. Vinokur and E. Burnstein, *Depolarization of attitudes in groups*, Journal of Personality and Social Psychology **36** (1978), 872.

[36] K. JS. Zollman, *The epistemic benefit of transient diversity*, Erkenntnis **72** (2010), 17–35.

Carlo Proietti,
Davide Chiarella

Conflict of Interest and the Principle of Total Evidence

1. *Introduction*

Recent literature suggests that most medical trials suffer from *conflict of interest* (henceforth, CoI), most notably financial sponsorship by pharmaceutical companies [24]. In this article, we address the unique challenge that evidence of CoI poses to the so-called *principle of total evidence* [2], which dictates that one ought to consider all sources of evidence when assessing the truth of a hypothesis.

To understand the nature of the challenge, it is instructive to first consider the much discussed—but conceptually simpler—case of evidence from randomized controlled trials (RCTs). RCTs are often treated as the gold standard of medical research [25]. The evidence-based-medicine movement holds that RCTs are at the top of the quality-of-evidence hierarchy. According to the "best evidence synthesis" view [27] only high-quality studies ought to be taken into account. Yet, RCTs have also been severely criticized for being subject to biases (*e.g.*, small size, inadequate blinding) that too often make them less than perfectly reliable [12, 32]. In this light, is it legitimate to consider other sources of evidence? In the literature, it has been argued [1, 22] that one ought to consider all sources of evidence, once information about reliability is taken into account, in line with the principle of total evidence.

Against this backdrop, how should one consider information *about CoI?* Similarly to the reliability case, we may regard the decision to ignore studies with CoI as an application of the best evidence view. Contrary to that case, however, it is unclear how one may demonstrate that it is useful to consider sources of evidence subject to CoI. The reason is that CoI has an *ambiguous influence* on reported results. In fact, available reviews suggest both that CoI raises the probability of biased estimates [7, 17] and that studies subject to CoI are more reliable in virtue of their better design/quality [19]. Intuitively, the two considerations pull in different directions, namely privileging evidence from RCTs with CoI vs discarding it. So the question arises, would one here, too, benefit from considering all sources of evidence?

Some authors [13,28] have recommended that conclusions from studies subject to CoI be down-adjusted, or "discounted". We see a fundamental problem with this proposal, namely that an unqualified discount, which does not take the dual role of CoI into account, would be epistemically unjust. Recently, Jonathan Fuller [8] has argued that evidence of studies subject to CoI can be appropriately taken into account—in a Bayesian way—rather than ignored. However, he proposes no concrete model to accomplish that.

Here, we provide a Bayesian model, which shows how one ought to take information on CoI into account without committing an epistemic injustice and in agreement with the principle of total evidence. Crucial to this task is that our model facilitates the understanding of how evidence affected by CoI can confirm true hypotheses and disconfirm false ones, precisely in virtue of representing all available information. Accordingly, we show that (A) studies subject to CoI can improve confirmation despite the ambiguous role of CoI, and that (B) information on CoI is not less relevant than other information, most notably, quality.

2. *Bias and the principle of total evidence*

Evidence-based medicine is the nowadays very popular view that there's a natural hierarchy in the quality of the evidence, and that one should prioritize evidence at the top and dismiss evidence at the bottom when assessing the support of medical studies for a hypothesis of interest. The strongest take on this view is that only the best available evidence should be considered ("best evidence synthesis", [27]), namely RCTs, and reviews and meta-analyses thereof.

RCTs are routinely used in the medical community to establish the efficacy of medical interventions with respect to pathologies of interest. Judgments of efficacy are based on outcome measures obtained in an RCT, typically the difference or the ratio between the effects of the intervention in the treatment group and in the control group. The goal of RCT is thus to estimate the causal effect of administering the treatment instead of the control, and ultimately, to warrant clinicians in using that intervention to treat patients suffering from those pathologies.

RCTs have two main virtues. By randomization over the test population, they are meant to eliminate confounding, namely the presence of factors other than the intervention making a difference to the outcome. Thanks to blind assignments of individuals to the treatment and the control (*e.g.*, placebo, alternative treatment) group, they are meant to eliminate biases where beliefs of the experimenters or the subjects may influence the outcome. This design ensures—in principle—a reliable estimate of the treatment's effect size. For this reason, RCTs are often considered the "gold standard" in medical research [25].

In practice, however, RCTs suffer from many limitations. These limitations are widely discussed and have led some to doubt the validity of much contemporary medical research [12]. Some limitations are statistical. For instance, small sample size increases the chances of unevenly distributed factors that bias the effect size estimate. Other limitations depend on faults in the experimental design. For instance, without proper blinding, selection bias induces experimenters/subjects to re-allocate. So, the question arises, is it legitimate to consider less-than-ideal evidence, in light of the biases of "best" evidence?

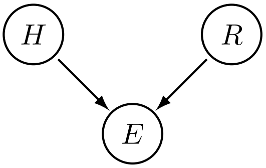
The question has been answered in the affirmative in the philosophical literature: evidence from less-than-ideal sources can improve confirmation if their reliability is properly accounted for [22]. The simplest scenario to illustrate this claim is a model over three variables, the hypothesis H, the evidence E, and the reliability of the evidential source R (Figure 1). Although the complete model includes a node denoting the reliability of the source, R, if the source is fully reliable, such that the prior of r equals 1, R drops out of the model, such that the confirmation of H (which is a function of its posterior probability relative to its prior) depends entirely on E:

$$P(h|e) = \frac{P(h)P(e|h)}{P(e)} = \frac{P(h)P(e|h)}{P(h)P(e|h) + P(\overline{h})P(e|\overline{h})}$$

By contrast, if the source of the evidence is not fully reliable, namely r has a prior smaller than 1, E has a weaker bearing on H, due to its dependence on R, once the latter is explicitly accounted for:

$$P(h|e) = \frac{P(h)P(r)P(e|hr) + P(h)P(\bar{r})P(e|h\bar{r})}{P(h)P(r)P(e|hr) + P(\overline{h})P(r)P(e|\overline{h}r) + P(h)P(\bar{r})P(e|h\bar{r}) + P(\overline{h})P(\bar{r})P(e|\overline{h}\bar{r})}.$$

Along similar lines, several methodologists advocate consideration of studies of both high- and low- quality when evaluating medical hypotheses. For instance, [30] and [29] have claimed that one may benefit from



$$H = \begin{cases} h: & \text{Clinically significant benefit.} \\ \bar{h}: & \text{No clinically significant benefit.} \end{cases}$$

$$R = \begin{cases} r: & \text{High reliability of the source.} \\ \bar{r}: & \text{Low reliability of the source.} \end{cases}$$

$$E = \begin{cases} e: & \text{Evidence of significant benefit.} \\ \bar{e}: & \text{No evidence of significant benefit.} \end{cases}$$

Figure 1

synthesizing evidence from studies randomly drawn from a mixture of studies with low risk of bias (in virtue of, respectively, allocation concealment or randomization) and studies with high risk of bias (for lack of concealment or randomization). Such proposals, like the simple model in Figure 1, promote the use of all available evidence over the neglect of part of it, even if it is of a lesser quality.

These considerations are in line with the principle of total evidence [2], which holds that, when assessing the credibility of hypotheses, we should endeavour to take into account all of the evidence at our disposal instead of just some proper part of it. The rationale behind the principle is that, since truth-conducive scientific inquiry is based on evidence, one ought not ignore any piece of evidence, as this would unnecessarily slow down the inquiry. That holds, note, even if the evidence in question is known to be potentially misleading. In that case, the principle requires that this piece of information—about the reliability of the evidence—be, too, taken into account to ensure maximal truth conduciveness whilst avoiding unjustified conclusions. In what follows, we shall endorse the principle of total evidence and argue that one ought to comply with it when accounting for the evidence, not only when the source is subject to a bias but also when it is subject to CoI.

3. *Conflict of interest*

An important but under-appreciated fact about RCTs is that most of them are subject to CoI [24]. CoI may be defined as "anything that may influence professional judgment", be that a financial interest by a pharmaceutical company, a personal interest, an academic interest, a political interest, or else [17, 2]. The most researched type of CoI is industry funding, and with good reason. There is strong evidence that industry funding is relevant to results aligning with the funders' interests [20, 21]. Moreover, it is plausible that other types of CoI—academic prestige, researcher allegiance, political/ideological beliefs, patient advocacy—may be relevant, too, yet there is no evidence that they actually are.[1] Therefore, our analysis shall focus on *financial* CoI rather than other types of CoI.

Assume that a large group of studies is partitioned into two subgroups depending on whether they are subject to CoI, which is in turn inferred from (reliable) reports on industry funding. To understand whether the difference between the two subgroups is significant, one might use a *meta-*

[1] The (scant) evidence on the matter actually suggests that other factors (*e.g.*, authors' affiliation, journal's impact factor) are not significant, once industry funding and study quality are controlled for [6].

regression as a diagnostic tool.[2] Assume that the effect size may be predicted as a linear function, for instance $f_i(x, y) = \alpha_i + \beta_i x + \gamma y$, of sampled effects x and study characteristics y, so as to minimize overall variance. By a meta-regression, one may test the statistical significance of additional covariates. For instance, one may treat CoI as a categorical (binary) covariate c in $f_i(x, c) = \alpha_i + \beta_i x + \gamma c$. Conditional on the value of c (*i.e.*, subgroup with/without CoI), the regression line may shift. To check if the shift is significant, one may test whether the null hypothesis $\gamma = 0$ can be rejected [5, 457]. By using statistical tools such as meta-regression, it has been observed that studies with CoI are 4 times as likely to produce positive outcomes [19]; 1.32 times as likely to report favourable efficacy results, and 1.87 times as likely to report favourable harms [20]; and generally more likely to produce favourable conclusions, if not favourable results [33].

Still, this observation alone does not establish which subgroup of studies is most trustworthy in producing more accurate estimates, as there are several possible explanations for the difference. The observation that industry-funded research is more likely to report significant beneficial effects is compatible with studies subject to CoI being *more*—rather than less—accurate. After all, industry-funded studies may have more power and better design, simply because they can exploit larger financial resources. This could fully explain the observed divergence in effects between the two subgroups. Indeed, while some studies report that industry funding is associated with poorer methodology [15,21], other studies report no difference in methodological quality between industry and non-industry funded research or even that industry funding is associated with higher quality [19].

What is striking, though, and in need of explanation, is the fact that the correlation between CoI and a difference in estimates is *robust*, even after controlling for a number of potential explanatory factors, such as methodological quality, statistical power, type of intervention or medical specialty [17], sample size, study design, country of primary authors [7], etc. This suggests that CoI may be a latent source of bias, even if the nature of the induced bias is hard to pin down exactly. Indeed, Lexchin [18] indicates that CoI can introduce biases in a subtle way—from selective outcome reporting (*e.g.*, standalone or repeated publication of study with favourable outcome), to poor design (*e.g.*, inappropriate choice of doses, dosing intervals, comparators), inadequate analysis (*e.g.*, “p-hacking”), and fraud (*e.g.*, data fabrication). Typically, one cannot verify whether all of the lat-

[2] On the difference between meta-regression and simple regression, see [11, §9.6.4].

ter factors are absent from a given study by directly inspecting the report of the study and of the data collected in the study. As a result, popular scores (*e.g.*, Risk-of-Bias), which are designed to evaluate the quality of the design of RCTs, are insensitive to such biases.

The above considerations should lead one to conclude that the effect of CoI is *ambiguous*—it can both improve quality *and* induce bias. How should one deal with this piece of information?

Ideally, if the sample size of studies without CoI were large enough, one could dispense with ambiguous evidence coming from potentially biased RCTs altogether. This is in line with the view that only the best available evidence should be considered. The problem with this view, however, is that large studies without CoI are often unavailable. For instance, phase III trials, although they may be conducted under the supervision of regulatory bodies, are typically sponsored by the pharmaceutical industry [23], which has a clear interest in demonstrating positive efficacy results. Drug approval thus has to rely on only few RCTs, which are normally all subject to CoI.

In light of this observation, some authors have recommended that conclusions from studies subject to CoI be down-adjusted, or "discounted". For instance, Stegenga [28] claims that "our confidence in medical interventions ought to be low, or at least much lower than is now the case". Ioannidis [13] advocates a "rational down-adjustment of effect sizes" and the use of "analytical methods that correct for anticipated inflation" (644). Fuller [8] maintains that, although multiple explanations for the association between CoI and divergent results remain possible, and although it is difficult to quantify the bias, "on net the plausible interpretations compel us to lower our confidence in therapies, at least qualitatively" (778). Yet, how to concretely go about discounting evidence subject to CoI is non-trivial. Given the association between CoI and bias, some [26,31] propose to add CoI to the categories that determine the quality score of a study. The problem with such proposals is that they may commit an *epistemic injustice.* In light of the evidence about the quality of individual studies subject to CoI, it would seem unjust to systematically punish studies with equal or superior design only because they belong to a "suspect" reference class. An unqualified discount of the evidence of studies subject to CoI tantamounts to neglecting its ambiguous role as both a promoter of biases and a preventer of them, and de facto to reducing CoI to a bias.

The lack of a concrete and well-motivated proposal on how, if at all, to discount evidence subject to CoI has important practical implications. For instance, meta-analyses are routinely used to improve effect size estimates by pooling together multiple RCTs and solve potential biases due to their individual (small) sample size. However, as observed by Roseman *et al.* [24], almost 70% of the RCTs included in meta-analyses are subject to CoI. Since CoI may decrease their reliability and skew their results, it poses a threat to the validity of the method, such that a prima facie more accurate estimate may hide biases induced by CoI, which actually lead to a less accurate estimate. In spite of this threat, the issue is not explicitly addressed by existing protocols on how to perform a meta-analysis.[3] As a result, current meta-analyses tend to omit any reference to CoI, let alone solve the possible problems due to them.

In sum, from the principle of total evidence, it follows that neglecting evidence from larger and more powerful studies is irrational, and so is neglecting evidence on the relevance of CoI to biased estimates of causal effects. At the same time, without a concrete model, and in light of CoI's ambiguous role, it's unclear whether discounting evidence subject to CoI is *justified* or how evidence from CoI-laden source is *confirmatory*. The rest of the paper is devoted to developing one such model. This, in turn, promises to be a first step towards addressing cognate problems, such as ensuring that meta-analyses actually improve effect size estimates by properly accounting for information on CoI.

4. *Fuller's proposal*

A recent proposal by Jonathan Fuller [8] on how to discount evidence subject to CoI shall serve as a useful starting point for our proposal.

The proposal is based on the distinction between two different types of evidence. There's *first-order evidence*—evidence E that bears directly on the probability of a hypothesis H—and *higher-order evidence*, which is evidence *about the evidence* E. In particular, in the latter category falls evidence E' of how the evidence E was generated. This sort of evidence is the product of "meta-research", which may be characterized as

> [...] an evolving scientific discipline that aims to evaluate and improve research practices. It includes thematic areas of methods, reporting, reproducibility, evaluation, and incentives [...] helping science progress faster by conducting scientific research on research itself. [14, 1-2]

In particular, meta-research evidence is "not evidence about a particular agent's reasoning; it concerns the public evidence from which many agents reason" [8, 774]. Based on this distinction, a particular study result is first-order evidence; whereas evidence on how the evidence is generated in the class of studies to which that particular study belongs is higher-order.

[3] For instance, the Cochrane Institute, which promotes the production of meta-analyses, does recommend methods of bias detection such as funnel plots and sensitivity analyses ([11, §10.4]), but provides no special recommendations on what to do with RCTs that report a CoI.

Fuller claims that one ought to incorporate meta-research evidence into a probabilistic (Bayesian) judgment on the hypothesis:

> Meta-evidence may be irrelevant to the bearing of E on H, but it is entirely relevant to our confidence in E and thus to our confidence in H. (774)

However, Fuller offers no concrete model to accompany his claim. The next section aims to remedy this deficiency.

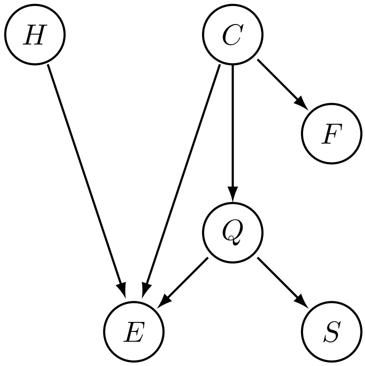
From Fuller's proposal, we inherit the attention to meta-research evidence and the view that both first- and higher-order evidence bear on the probability of the hypothesis. At the same time, Fuller invites an equivocation between the "directness" of the evidence and its "order" [8, 770], which notions we prefer to keep distinct. Moreover, we do not share his intuitions on the appropriate (Bayesian) way to model this scenario. He maintains that E' is relevant to H because E fails to screen off [8, 776], or render conditionally independent, E' and H. However, this suggests that there is a missing edge between E' and H relative to either $E' \longleftarrow E \longrightarrow H$ or $E' \longrightarrow E \longrightarrow H$, both of which structures contradict not only Fuller's statement that there is no direct dependence between meta-research and hypothesis (*cf.* quote) but also the widely received view that the reliability of the evidence-gathering method bears on the hypothesis via a collider structure, $H \longrightarrow E \longleftarrow E'$ (*cf.* Figure 1, with E' in the place of R). The next section shall introduce a fully fledged Bayesian model that explains how evidence of RCTs subject to CoI may aid confirmation without being subject to the aforementioned problems.

5. *A Bayesian model*

We now introduce a Bayesian model, which we motivate and set up in agreement with the principle of total evidence. Ultimately, the model shall serve to prove two main theses, namely that (A) information on CoI can improve confirmation despite the inherent uncertainty on CoI's opposite effects, and irrespective of the quality of a study, and that (B) it is generally false that CoI has a smaller confirmatory weight than quality, and thus it is unjustified to preferentially neglect CoI over quality.

5.1. *Variables and dependencies*

For simplicity, all of the variables in the model are binary (Figure 2). The hypothesis of interest depends on many sources of evidence. The first obvious difference between these sources is that some are *direct* relative to the hypothesis, H, and others are *indirect*, by which we mean that the former are graphically adjacent to H, and the latter influence H via other nodes. In particular, direct evidence for H are reported effects, E. Indirect evidence for H are financial conflict of interest C, industry funding F,



$$H = \begin{cases} h: & \text{Clinically significant benefit.} \\ \bar{h}: & \text{No clinically significant benefit.} \end{cases}$$

$$C = \begin{cases} c: & \text{Financial conflict of interest.} \\ \bar{c}: & \text{No financial conflict of interest.} \end{cases}$$

$$F = \begin{cases} p: & \text{Reported industry funding.} \\ \bar{p}: & \text{No reported industry funding.} \end{cases}$$

$$Q = \begin{cases} q: & \text{High study quality.} \\ \bar{q}: & \text{Low study quality.} \end{cases}$$

$$S = \begin{cases} s: & \text{High quality score.} \\ \bar{s}: & \text{Low quality score.} \end{cases}$$

$$E = \begin{cases} e: & \text{Evidence of significant benefit.} \\ \bar{e}: & \text{No evidence of significant benefit.} \end{cases}$$

FIGURE 2

study quality Q, and quality scores S, because all of them influence H via E. This aligns with modelling the reliability node, C, by a collider structure, $H \longrightarrow E \longleftarrow C$, where E connects H and C, contrary to Fuller's suggestion that H is a descendant of C—which he labels E'—and that E mediates C's influence but fails to screen it off from H.

A second difference between the sources of evidence is that they are modelled differently: some are modelled as *nodes* (*cf.* [1, chapter 4]), whose values are either directly observed, such as the reported effects E, or informed by proxies. For instance, one may use information in the individual studies, scored by quality assessment tools [28, chapter 7], S, as a proxy for study validity, or quality, Q. And one may use funding reports F as a proxy for CoI [20], C, indicating whether CoI is in fact present or absent. By contrast, other sources of evidence are conceptualized as *parameters*. Crucially, some such parameters, for instance the likelihood of bias induced by CoI, are constrained by evidence of meta-epidemiological data from appropriate reference classes of studies where as much information as possible is available on factors to which the likelihoods are sensitive (*e.g.*, truth of the hypothesis, quality of the study).

Now we're in a position to better characterize the distinction between first and higher-order evidence for H. First-order evidence for H are facts discovered by observation, viz. E, F, S. Higher-order evidence for H are constraints inferred from meta-research (*e.g.*, meta-regression), such as the (ceteris paribus) residual influence of CoI on the evidence. For instance, S is first-order evidence for H because quality scores aggregate observable characteristics of quality. All criteria in each quality tool are *testable on individual studies independently*. By contrast, the robust dependence of the evidence on CoI is higher-order evidence for H because it denotes a biasing influence of CoI on the evidence via characteristics, such as ghostwriting, fraud, etc., which are unobservable at the individual level, and thus to which the scores are insensitive [18], and are *only testable at the population level.* Note that the distinction between first- and higher-order evidence is orthogonal to that between direct and indirect evidence. For instance, C is indirect evidence for H but need *not* be higher-order wrt H.

In this respect, evidence of the residual influence of CoI is analogous to evidence of biases, which may only be assessed at the population level. One example is publication bias [4], namely the overestimation of the effect size due to the fact that insignificant results are less likely to be published or even submitted for publication. The bias becomes visible by the use of so-called funnel plots, which represent the estimates of multiple studies against their size. Without bias, one should expect a symmetric distribution, with estimates of larger studies (with higher precision) being near the average and estimates of smaller studies (with lower precision) being spread evenly on both sides of the distribution. The bias induces an uneven distribution of the estimates of small studies, such that there are more observations on the right side than on the left side. Publication bias, thus, is detectable only at the level of populations of studies.[4] In what sense, then, is CoI analogous to publication bias, and in what sense is it different?

Unlike publication bias, CoI is not a bias. Rather, it is a possible *cause* of biases, some of which are detectable at the individual level and some at the population level. Like evidence of publication bias, however, evidence of CoI is a population-level property, something over and above evidence that may be gathered by looking at individual studies. These features of CoI are directly reflected in our model. In it, CoI is not a bias but a "cause" of biases, which may both promote and hamper confirmation. This is rendered in the model by having C influence E along *two* paths, one where it promotes quality by preventing biases detectable at the individual level, and one where it promotes biases undetectable at the individual level—although they may be detectable at the population level, if suitable proxies are available.[5] Therefore, in our model, evidence subject to CoI may or may not need discounting, depending on higher-order evidence of

[4] One way to correct for the bias is to re-estimate the effect size not based on the actual, unevenly distributed population, but on the counterfactual, evenly distributed one, to which the (putatively) missing studies have been added.

[5] For simplicity, our model leaves implicit the latter kind of biases as well as any proxy of them. If one were to explicitly model those nodes, they would be graphically analogous to Q and S, only located on, or departing from, the direct edge $C \longrightarrow E$.

the strength of CoI on either path, and on first-order evidence, if any, on whether CoI generates biases along the two paths.

5.2. *Probabilistic constraints*

The following constraints, and all of our below results, are relative to the case of Bayesian confirmation of a hypothesis by the evidence from a single study—possibly subject to CoI—in the absence of further evidence. Some of the constraints are quantitative, some are qualitative. Among the former, some are directly informed by meta-research evidence. In the conclusion, we will briefly touch on the issues arising in the case of confirmation by evidence from multiple studies.

$$P(h) \in [0.001, 0.01] \tag{5.1}$$

$$P(c) \in [0.7, 1] \tag{5.2}$$

$$P(f|c) = 0.9 \quad P(f|\bar{c}) = 0.1 \tag{5.3}$$

$$P(q|c) = [0.5, 0.9] \quad P(q|\bar{c}) = 0.5 \tag{5.4}$$

$$P(s|q) = 0.9 \quad P(s|\bar{q}) \in [0.3, 0.7] \tag{5.5}$$

$$P(e|c)/P(e|\bar{c}) \in [1.3, 4] \tag{5.6}$$

$$P(e|h) > P(e|\bar{h}) \tag{5.7}$$

$$1 \geq P(e|hqc) > P(e|h\bar{q}c) \quad P(e|hq\bar{c}) > P(e|h\bar{q}\bar{c}) \tag{5.8}$$

$$P(e|\bar{h}qc) < P(e|\bar{h}\bar{q}c) \quad 0 \leq P(e|\bar{h}q\bar{c}) < P(e|\bar{h}\bar{q}\bar{c}) \tag{5.9}$$

$$0 \leq P(e|\bar{h}q\bar{c}) < P(e|\bar{h}qc) \quad P(e|\bar{h}\bar{q}\bar{c}) < P(e|\bar{h}\bar{q}c) \tag{5.10}$$

$$P(e|hq\bar{c}) < P(e|hqc) \leq 1 \quad P(e|h\bar{q}\bar{c}) < P(e|h\bar{q}c) \tag{5.11}$$

Let us examine the above constraints one by one, starting with the quantitative constraints (5.1) to (5.6). (5.1) says that the prior of a (novel) hypothesis being true is low. We assume that major discoveries of treatments with huge benefits have been made already (penicillin, etc.) and still-to-be-made discoveries are fewer and their effect sizes are smaller. Whereas critics such as Ioannidis and Stegenga will be more conservative on the prior of novel findings on clinically relevant benefits (say, 1/1000), the pharmaceutical industry will arguably be less conservative (say, 1/100). (5.2) says the probability of CoI in the overall population of studies is over 70%. This number is suggested by Roseman *et al.* [24], after a count of the declarations of CoI in studies included in meta-analyses. This figure, in turn, is arguably representative of the percentage of studies subject to CoI in the overall population of studies, whether or not they are included in meta-analyses. (5.3) says that financial CoI is very likely to induce an industry funding report, and its absence is unlikely to induce an industry funding report. Notice that, although funding can come from both no-profit and for-profit organizations, also at the same time [17, 3], for-profit organizations have arguably more money to invest and more incentive to invest in studies where they can skew the result in their favour. As such, industry funding reports are strong evidence of financial CoI, even in the presence of public funding. (5.4) reflects that quality is often considered higher (between 30% to 40% higher) in industry-sponsored studies, and sometimes considered the same as in publicly funded studies [19, 10]. It is seldom considered inferior [15, 21]. These assessments depend on the quality measure chosen to test for differences, the field of studies considered in the test, etc. By contrast, we assume that there is no distinctive quality expectation if the study is publicly funded. The (scant) research on the topic suggests that publicly funded studies tend to, if anything, be at a higher risk of bias [9]. (5.5) says that the likelihood of a high score—by whatever quality measure—given high quality is arguably large, but the likelihood of a high score given low quality tends to be sensitive to the chosen measure, given that each measure is unable to spot at least some internal validity problem, and some measures are worse than others, at least according to Stegenga [28]. It is difficult to extract intervals—let alone reliable ones—from Stegenga's sources, viz. [10] and [16]. For the sake of illustration, we pick $[0.3, 0.7]$ as our chosen interval, but nothing in our results hinges on this choice. Finally, the crucial quantitative piece of information for our subsequent argument is given by (5.6), which says that CoI is expected to render favourable results from 1.3 times to 4 times more likely, depending on the survey [19, 20]. A likelihood ratio larger than 1 was verified to be robust to controls for design features and statistical power, among other characteristics [7, 17].

Next follow a number of qualitative constraints. (5.7) says that (ceteris paribus) the truth of the hypothesis makes positive evidence e more likely. (5.8) and (5.9) say that (ceteris paribus) lowering the quality of the study design makes the study worse at tracking the truth. If $\bar{h}$ is the case, then a low quality study will make e more likely than a higher quality one. If h is the case, then a low quality study will make e less likely than a higher quality one. (5.10) and (5.11) say that (ceteris paribus) CoI makes e more likely. Note that combining (5.8) and (5.11) gives

$$1 = P(e|hqc) > P(e|h\bar{q}c), P(e|hq\bar{c}) > P(e|h\bar{q}\bar{c}). \tag{5.12}$$

That is, when h is the case, CoI and high quality raise the probability of e. Moreover, combining (5.9) and (5.10) gives

$$0 = P(e|\bar{h}q\bar{c}) < P(e|\bar{h}qc), P(e|\bar{h}\bar{q}\bar{c}) < P(e|\bar{h}\bar{q}c). \tag{5.13}$$

That is, when $\bar{h}$ is the case, CoI and *low* quality raise the probability of e. By inspecting (5.12) and (5.13), one may see that (5.8) to (5.11) entail that c raises the probability of e, no matter the values of H and Q. In particular, CoI increases the probability of e both if H is true—because CoI raises quality, which makes it more likely to get evidence confirming h—and if H is false—because CoI decreases the probability of $\bar{e}$, which makes it less likely to get evidence disconfirming h. In sum, CoI has both a positive and a negative influence on truth-tracking. Which of the two prevails depends on the strengths of its effects on e and q.

5.3. *Meta-research and the dual role of conflict of interest*

Even if it is accepted by a rational agent who endorses (5.8) to (5.11) that CoI has both a positive and a negative influence on truth-tracking, neither those qualitative constraints alone, nor the addition of the quantitative ones in (5.1) to (5.5)—but, crucially, not (5.6)—would force them to conclude that C has a *significant* biasing influence on E, which in turn translates into a *non-negligible* parameter of the influence along the direct edge $C \longrightarrow E$. In fact, while (5.10) and (5.11) maintain that C is (ceteris paribus) relevant to E, they are consistent with this influence being extremely small given Q, to the point that the existence of a direct edge $C \longrightarrow E$ may be neglected without detriment.

Now, let us consider adding (5.6) to one's total evidence. To begin with, note that the ratio in (5.6) may be unpacked as follows:

$$\frac{P(e|c)}{P(e|\bar{c})} = \frac{P(ec) \cdot P(\bar{c})}{P(e\bar{c}) \cdot P(c)} = \frac{P(\bar{c})}{P(c)} \cdot \frac{P(hecq) + P(hec\bar{q}) + P(\bar{h}ecq) + P(\bar{h}ec\bar{q})}{P(he\bar{c}q) + P(he\bar{c}\bar{q}) + P(\bar{h}e\bar{c}q) + P(\bar{h}e\bar{c}\bar{q})}$$
$$= \frac{P(e|hcq) \cdot P(h) \cdot P(q|c) + P(e|hc\bar{q}) \cdot P(h) \cdot P(\bar{q}|c) + P(e|\bar{h}cq) \cdot P(\bar{h}) \cdot P(q|c) + P(e|\bar{h}c\bar{q}) \cdot P(\bar{h}) \cdot P(\bar{q}|c)}{P(e|h\bar{c}q) \cdot P(h) \cdot P(q|\bar{c}) + P(e|h\bar{c}\bar{q}) \cdot P(h) \cdot P(\bar{q}|\bar{c}) + P(e|\bar{h}\bar{c}q) \cdot P(\bar{h}) \cdot P(q|\bar{c}) + P(e|\bar{h}\bar{c}\bar{q}) \cdot P(\bar{h}) \cdot P(\bar{q}|\bar{c})}$$

Given that $P(h)$ is by assumption very small (5.1), that the first two terms in numerator and denominator are negligible, and that $P(\bar{h})$ in the other terms can be factored out and simplified, the expression approximates to

$$\frac{P(e|c)}{P(e|\bar{c})} \approx \frac{P(e|\bar{h}cq) \cdot P(q|c) + P(e|\bar{h}c\bar{q}) \cdot P(\bar{q}|c)}{P(e|\bar{h}\bar{c}q) \cdot P(q|\bar{c}) + P(e|\bar{h}\bar{c}\bar{q}) \cdot P(\bar{q}|\bar{c})} = \frac{P(e|\bar{h}cq) \cdot P(q|c) + P(e|\bar{h}c\bar{q}) \cdot P(\bar{q}|c)}{[P(e|\bar{h}\bar{c}q) + P(e|\bar{h}\bar{c}\bar{q})]/2}.$$

A conservative yet rational agent, who believes that C is almost irrelevant to E given Q, will judge the terms $\frac{P(e|\bar{h}cq)}{P(e|\bar{h}\bar{c}q)}$ and $\frac{P(e|\bar{h}c\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})}$ close to 1, such that the ratio will ultimately be driven by $P(q|c)$ and $P(q|\bar{c})$. In that case,

$$\begin{aligned}\frac{P(e|c)}{P(e|\bar{c})} &\approx \frac{P(e|\bar{h}cq)\cdot P(q|c)+P(e|\bar{h}c\bar{q})\cdot P(\bar{q}|c)}{[P(e|\bar{h}\bar{c}q)+P(e|\bar{h}\bar{c}\bar{q})]/2}\\ &= \frac{\frac{P(e|\bar{h}cq)}{2}+P(e|\bar{h}c\bar{q})\cdot P(\bar{q}|c)}{[P(e|\bar{h}\bar{c}q)+P(e|\bar{h}\bar{c}\bar{q})]/2}+\frac{(0.9-P(c|q))\cdot P(e|\bar{h}cq)}{[P(e|\bar{h}\bar{c}q)+P(e|\bar{h}\bar{c}\bar{q})]/2}\\ &\leq \frac{[P(e|\bar{h}cq)+P(e|\bar{h}c\bar{q})]/2}{[P(e|\bar{h}\bar{c}q)+P(e|\bar{h}\bar{c}\bar{q})]/2}+\frac{0.4\cdot P(e|\bar{h}cq)}{[P(e|\bar{h}\bar{c}q)+P(e|\bar{h}\bar{c}\bar{q})]/2}\\ &\leq 1+0.8\cdot\frac{P(e|\bar{h}cq)}{P(e|\bar{h}\bar{c}q)+P(e|\bar{h}\bar{c}\bar{q})} < 1+0.8\cdot\frac{P(e|\bar{h}cq)}{P(e|\bar{h}\bar{c}q)} = 1.8.\end{aligned}$$

That is, a rational agent, who is also a strong supporter of the confirmatory value of C via its positive effect on Q will assign values, which in the most favourable case, $P(q|c) = 0.9$, return an upper bound for interval of the likelihood ratio equal to 1.8.

At the same time, coherence requires avoiding probability assignments violating (5.6), once the latter piece of evidence becomes known. A rational agent will thus have to accommodate their beliefs in light of (5.6) to ensure that the range of possible values of the ratio $\frac{P(e|c)}{P(e|\bar{c})}$ *contains the interval* $[1.3, 4]$. The only probabilities, which have not yet been pinned down, are those of E conditional on its parents H, Q, C. Since $P(h)$ being very small makes almost irrelevant probabilities of e conditional on h, $P(e|h...)$, the most plausible way for a rational agent to adjust their beliefs is to increase the value of either the ratio $P(e|\bar{h}cq)/P(e|\bar{h}\bar{c}q)$, or the ratio $P(e|\bar{h}c\bar{q})/P(e|\bar{h}\bar{c}\bar{q})$, or both. This boils down to acknowledging that C has a relevant effect on the study's outcome e, conditional on its quality Q. In other words, the principle of total evidence dictates that a conservative yet rational agent assign to CoI a non-redundant role for confirmation, given quality.

5.4. *Conflict of interest and Bayesian confirmation*

Let us now come to the illustration of how the model supports our main claims, starting with (A): it is unjustified to generally neglect CoI despite its ambiguity, because CoI can make a difference —whether positive or negative— to the confirmation of the hypothesis.

Theorem 5.1 (Confirmation by CoI-laden study). *A CoI-laden study is confirmatory if the relevant Bayes factor* > 1*:*

$$\text{sign}(P(h|ec)-P(h))=\text{sign}\left(\frac{P(e|hcq)\cdot P(q|c)+P(e|hc\bar{q})\cdot P(\bar{q}|c)}{P(e|\bar{h}cq)\cdot P(q|c)+P(e|\bar{h}c\bar{q})\cdot P(\bar{q}|c)}-1\right).$$

Theorem 5.1 says that CoI-laden studies may (dis)confirm. This supports the view that such studies should not be ignored but suitably discounted.

Theorem 5.2 (Confirmation by CoI-laden study, irrespective of quality). *A CoI-laden study is more confirmatory for* $H = h$ *iff* $P(h|e\bar{c}) < P(h|e)$*:*

$$\text{sign}(P(h|ec) - P(h|e\bar{c})) = \text{sign}(P(h|ec) - P(h|e)) = \text{sign}\left(\frac{P(e|hc)}{P(e|\bar{h}c)} - \frac{P(e|h\bar{c})}{P(e|\bar{h}\bar{c})}\right)$$

$$= \text{sign}\left(\frac{P(e|hcq) \cdot P(q|c) + P(e|hc\bar{q}) \cdot P(\bar{q}|c)}{P(e|\bar{h}cq) \cdot P(q|c) + P(e|\bar{h}c\bar{q}) \cdot P(\bar{q}|c)} - \frac{P(e|h\bar{c}q) \cdot P(q|\bar{c}) + P(e|h\bar{c}\bar{q}) \cdot P(\bar{q}|\bar{c})}{P(e|\bar{h}\bar{c}q) \cdot P(q|\bar{c}) + P(e|\bar{h}\bar{c}\bar{q}) \cdot P(\bar{q}|\bar{c})}\right).$$

In particular, Theorem 5.2 entails that CoI can not only decrease but also *increase* confirmation, namely if c is a good predictor of $H = h$, that is, if c makes e much more likely given h, and only slightly more likely given $\bar{h}$.

Next, let us turn to prove (B): it is unjustified to generally (or necessarily, or a priorily) neglect CoI vis-à-vis quality.

Theorem 5.3 (Confirmation by low-quality & no-CoI vs CoI & high-quality).

$$\begin{aligned}\text{sign}(P(h|e\bar{c}\bar{q}) - P(h|ecq)) &= \text{sign}\left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right) \\ &= \text{sign}\left(\frac{P(e|\bar{h}cq)}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hcq)}{P(e|h\bar{c}\bar{q})}\right).\end{aligned}$$

Theorem 5.3 shows that neither CoI given high quality nor quality given absence of CoI is generally less relevant to confirmation. Thus, one cannot conclude that it is more justifiable to neglect one at the expenses of the other on the assumption that this would result in a smaller violation of the principle of total evidence. Loosely speaking, whether an industry-sponsored randomized study is more confirmatory than an independent non-randomized study depends on concrete circumstances.

Theorem 5.4 (Confirmation by low-quality vs CoI).

$$\begin{aligned}&\text{sign}(P(h|e\bar{q}) - P(h|ec) \\ &= \text{sign}\Bigg(\left(\frac{P(e|hc\bar{q})}{P(e|\bar{h}c\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right) P(e|\bar{h}c\bar{q})\, P(e|\bar{h}cq)\, P(c)^2\, P(q|c)\, P(\bar{q}|c) \\ &\qquad + \left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right) P(e|\bar{h}\bar{c}\bar{q})\, P(e|\bar{h}cq)\, P(c)\, P(\bar{c})\, P(q|c)\, P(\bar{q}|\bar{c}) \\ &\qquad + \left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hc\bar{q})}{P(e|\bar{h}c\bar{q})}\right) P(e|\bar{h}\bar{c}\bar{q})\, P(e|\bar{h}c\bar{q})\, P(c)\, P(\bar{c})\, P(\bar{q}|c)\, P(\bar{q}|\bar{c})\Bigg).\end{aligned}$$

Theorem 5.4 considers the relative confirmatory strength of low-quality studies vs CoI-laden studies. It shows that the comparison between the two is not straightforward, as it is sensitive to a large number of parameters. From this, it follows that is not a priori justified to exclude from consideration either ones between low-quality studies and CoI-laden studies.

Of course, to disentangle CoI's truth-tracking contribution, namely the one via Q, from the "truth-diverting" contribution, namely the one that introduces biases via E directly, one needs (plausible) numerical assignments for all of the probabilities above. If one can precisely assign such values, one can use the theorem to verify which contribution prevails, by looking at the sign of $(P(h|eq)-P(h))-(P(h|ec)-P(h)) = P(h|eq)-P(h|ec)$. If we employ a notion of rationality where imprecise probabilities are allowed, however, there is no guarantee that the sign is conclusively decidable. If meta-research does not provide sufficient information to pin down all probabilities, the model may fail to eliminate the uncertainty on whether there is confirmation or not.

At the same time, note that an ideal Bayesian agent assigns exact values to all of the probabilities in the above theorem, which makes it in principle possible to use our model to definitely determine whether CoI confirms or disconfirms on any given occasion. Moreover, even if—more realistically—meta-research does not justify precise assignments, it may still suffice to constrain the probabilities to intervals, which are narrow enough, such that the value of that difference does not span both positive and negative values, and the only uncertainty left concerns the amount of confirmation but not its sign. In this sense, the model should be viewed as a roadmap, which allows (non-ideal) agents to collect as much information as needed for the model to return acceptable answers, given the level of accuracy required by the problem at hand.

6. *Conclusion*

Existing literature on Bayesian confirmation supports the view that less-than-ideal evidence—from, say, medical studies—provides confirmatory benefits if a reliability-based discount is introduced to account for the likelihood of the evidence being produced not only by the truth of the hypothesis but also by underlying biases. This is in line with the principle of total evidence, which recommends that one takes into account all available evidence when determining probabilities.

At the same time, it has been noted that medical studies are subject not only to biases but also to conflicts of interest. It is unclear, however, whether evidence obtained from studies subject to a conflict of interest, too, should be discounted, given that conflict of interest has an ambiguous role. On the one hand, in fact, it promotes larger studies with better designs, which improves their accuracy by making certain biases less likely. On the other hand, it introduces subtle and difficult-to-detect biases, which may tend to skew the results, by for instance overestimating effect sizes. Without a concrete model, however, it is difficult to assess the bearing of conflict of interest on the hypothesis under scrutiny, and more generally, the benefit of complying with the principle of total evidence by taking evidence of conflict of interest into account.

This paper has addressed the issue by providing a Bayesian model, which disambiguates the dual role of conflict of interest as a promoter of bias and of quality along distinct paths, and interprets its influences along the two paths as higher-order constraints obtained from meta-research. The analysis of the model shows that information on conflict of interest can benefit confirmation. More precisely, it is unjustified to neglect conflict of interest, whether or not information on quality is available, or to prefer the neglect of conflict of interest vis-à-vis quality. Together, these results vindicate the endorsement of the principle of total evidence in the presence of conflict of interest.

Two important issues have not been addressed here. First, our model only considers Bayesian updates by a single study. It can be extended to updates by evidence from multiple studies. In that case, however, the choice of the model will depend on the source(s) of conflict of interest behind the studies, more precisely on whether each study is subject to a different conflict of interest, or all studies are subject to the same conflict of interest or, more realistically, on where the studies are located on an interval between these two extremes. The set up and the analysis of this extended model raises technical issues, which cannot be reviewed here (the interested reader is referred to [3]).

Second, our model uses binary variables. Only considering the significance of a result is, of course, a gross simplification. Public health decisions depend on numerical estimates of the effect size. The impact of conflict of interest on the veracity of a study depends crucially on the observed effect size. Among significant studies subject to conflict of interest, some may show a larger effect than others, and this is certainly relevant to the bearing of the study on the estimate, and to how the estimate should be revised in light of the evidence. For a Bayesian model to inform such estimates, one must specify how a continuous prior of an effect size (characterized by, say, a given mean and variance) should be updated given the evidence (an empirical distribution with another mean and variance) under the assumed constraints. To this end, however, analytic results are not sufficient and one must resort to numerical simulations (see, *e.g.*, [30] and [29]). Such simulations, in turn, could be instrumental to revising estimates by not only single studies but also collections of such studies, as is typical done by meta-analyses.

Jointly, these two issues motivate a study of how to devise a meta-analytic revision of the effect size (in the presence of conflicts of interest) in a Bayesian framework. We leave this study to future research.

Acknowledgements. Lorenzo Casini acknowledges funding from the Italian Ministry of Scientific Research and Education (MIUR) under the grants 201743F9YE and 20177FX2A7 and under the PRO3 grant for the project "Understanding public data: experts, decisions, epistemic values". Jürgen Landes gratefully acknowledges NextGenerationEU funding for the project "Practical Reasoning for Human-Centred Artificial Intelligence" as well as funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the grant 528031869. We thank the audience of the workshop *Reasoning with Imperfect Information in Social Settings*, SNS Pisa, 28 October 2023, where this paper was presented. A special thank goes to Jan Sprenger and Mattia Andreoletti for helpful comments and discussions.

Proofs

Proof of Theorem 5.1 (Confirmation by CoI-laden study).

$$\begin{aligned}
\text{sign}(P(h|ec) - P(h)) &= \text{sign}(P(hec) - P(h) \cdot P(ec)) \\
&= \text{sign}(P(hec) - P(h) \cdot [P(hec) + P(\bar{h}ec)]) \\
&= \text{sign}(P(\bar{h}) \cdot P(hec) - P(h) \cdot P(\bar{h}ec)]) \\
&= \text{sign}\Big(\frac{P(hecq) + P(hec\bar{q})}{P(\bar{h}ecq) + P(\bar{h}ec\bar{q})} - \frac{P(h)}{P(\bar{h})}\Big) \\
&= \text{sign}\Big(\frac{P(e|hcq) \cdot P(cq) + P(e|hc\bar{q}) \cdot P(c\bar{q})}{P(e|\bar{h}cq) \cdot P(cq) + P(e|\bar{h}c\bar{q}) \cdot P(c\bar{q})} - 1\Big) \\
&= \text{sign}\Big(\frac{P(e|hcq) \cdot P(q|c) + P(e|hc\bar{q}) \cdot P(\bar{q}|c)}{P(e|\bar{h}cq) \cdot P(q|c) + P(e|\bar{h}c\bar{q}) \cdot P(\bar{q}|c)} - 1\Big).
\end{aligned}$$

□

Proof of Theorem 5.2 (Confirmation by CoI-laden study, irrespective of quality).

$$\begin{aligned}
&\text{sign}(P(h|e\bar{c}) - P(h|ec)) = \text{sign}([P(he\bar{c}) \cdot P(\bar{h}ec)] - [P(hec) \cdot P(\bar{h}e\bar{c})]) \\
&= \text{sign}([(P(he\bar{c}q) + P(he\bar{c}\bar{q})) \cdot (P(\bar{h}ecq) + P(\bar{h}ec\bar{q}))] \\
&\qquad - [(P(hecq) + P(hec\bar{q})) \cdot (P(\bar{h}e\bar{c}q) + P(\bar{h}e\bar{c}\bar{q}))]) \\
&= \text{sign}([P(h) \cdot (P(e|h\bar{c}q) \cdot P(\bar{c}q) + P(e|h\bar{c}\bar{q}) \cdot P(\bar{c}\bar{q})) \cdot P(\bar{h}) \cdot (P(e|\bar{h}cq) \cdot P(cq) \\
&\qquad + P(e|\bar{h}c\bar{q}) \cdot P(c\bar{q}))] \\
&\qquad - [P(h) \cdot (P(e|hcq) \cdot P(cq) + P(e|hc\bar{q}) \cdot P(c\bar{q})) \cdot P(\bar{h}) \cdot (P(e|\bar{h}\bar{c}q) \cdot P(\bar{c}q) \\
&\qquad\quad + P(e|\bar{h}\bar{c}\bar{q}) \cdot P(\bar{c}\bar{q}))]) \\
&= \text{sign}([(P(e|h\bar{c}q) \cdot P(q|\bar{c}) + P(e|h\bar{c}\bar{q}) \cdot P(\bar{q}|\bar{c})) \cdot (P(e|\bar{h}cq) \cdot P(q|c) + P(e|\bar{h}c\bar{q}) \cdot P(\bar{q}|c))] \\
&\qquad - [(P(e|hcq) \cdot P(q|c) + P(e|hc\bar{q}) \cdot P(\bar{q}|c)) \cdot (P(e|\bar{h}\bar{c}q) \cdot P(q|\bar{c}) + P(e|\bar{h}\bar{c}\bar{q}) \cdot P(\bar{q}|\bar{c}))]) \\
&= \text{sign}\Big(\frac{P(e|h\bar{c}q) \cdot P(q|\bar{c}) + P(e|h\bar{c}\bar{q}) \cdot P(\bar{q}|\bar{c})}{P(e|\bar{h}\bar{c}q) \cdot P(q|\bar{c}) + P(e|\bar{h}\bar{c}\bar{q}) \cdot P(\bar{q}|\bar{c})} - \frac{P(e|hcq) \cdot P(q|c) + P(e|hc\bar{q}) \cdot P(\bar{q}|c)}{P(e|\bar{h}cq) \cdot P(q|c) + P(e|\bar{h}c\bar{q}) \cdot P(\bar{q}|c)}\Big).
\end{aligned}$$

□

Proof of Theorem 5.3 (Confirmation by low-quality & no-CoI vs CoI & high-quality)

$$
\begin{aligned}
&\text{sign}(P(h|e\bar{c}\bar{q}) - P(h|ecq)) = \text{sign}(P(he\bar{c}\bar{q}) \cdot P(ecq) - P(he\bar{c}\bar{q}) \cdot P(e\bar{c}\bar{q}))\\
&\quad = \text{sign}(P(he\bar{c}\bar{q}) \cdot [P(hecq) + P(\bar{h}ecq)] - P(hecq) \cdot [P(he\bar{c}\bar{q}) + P(\bar{h}e\bar{c}\bar{q})])\\
&\quad = \text{sign}(P(he\bar{c}\bar{q}) \cdot P(\bar{h}ecq) - P(hecq) \cdot P(\bar{h}e\bar{c}\bar{q}))\\
&\quad = \text{sign}(P(h) \cdot P(\bar{h}) \cdot P(cq) \cdot P(\bar{c}\bar{q}) \cdot [P(e|h\bar{c}\bar{q}) \cdot P(e|\bar{h}cq) - P(e|hcq) \cdot P(e|\bar{h}\bar{c}\bar{q})])\\
&\quad = \text{sign}(P(e|h\bar{c}\bar{q}) \cdot P(e|\bar{h}cq) - P(e|hcq) \cdot P(e|\bar{h}\bar{c}\bar{q}))\\
&\quad = \text{sign}\left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right). \qquad \square
\end{aligned}
$$

Proof of Theorem 5.4 (Confirmation by low-quality vs CoI).

$$
\begin{aligned}
&\text{sign}(P(h|e\bar{q}) - P(h|ec)) = \text{sign}(P(he\bar{q}) \cdot [P(hecq) + P(hec\bar{q}) + P(\bar{h}ecq) + P(\bar{h}ec\bar{q})]\\
&\qquad - P(hec) \cdot [P(hec\bar{q}) + P(he\bar{c}\bar{q}) + P(\bar{h}ec\bar{q}) + P(\bar{h}e\bar{c}\bar{q})])\\
&\quad = \text{sign}([P(hec\bar{q}) + P(he\bar{c}\bar{q})] \cdot [P(hecq) + P(hec\bar{q}) + P(\bar{h}ecq) + P(\bar{h}ec\bar{q})]\\
&\qquad - [P(hecq) + P(hec\bar{q})] \cdot [P(hec\bar{q}) + P(he\bar{c}\bar{q}) + P(\bar{h}ec\bar{q}) + P(\bar{h}e\bar{c}\bar{q})])\\
&\quad = \text{sign}(P(hec\bar{q}) \cdot P(\bar{h}ecq) + P(he\bar{c}\bar{q}) \cdot P(\bar{h}ecq) + P(he\bar{c}\bar{q}) \cdot P(\bar{h}ec\bar{q})\\
&\qquad - [P(hecq) \cdot P(\bar{h}ec\bar{q}) + P(hecq) \cdot P(\bar{h}e\bar{c}\bar{q}) + P(hec\bar{q}) \cdot P(\bar{h}e\bar{c}\bar{q})])\\
&\quad = \text{sign}([P(e|hc\bar{q}) \cdot P(e|\bar{h}cq) - P(e|hcq) \cdot P(e|\bar{h}c\bar{q})] \cdot P(h) \cdot P(\bar{h}) \cdot P(c\bar{q}) \cdot P(cq)\\
&\qquad + [P(e|h\bar{c}\bar{q}) \cdot P(e|\bar{h}cq) - P(e|hcq) \cdot P(e|\bar{h}\bar{c}\bar{q})] \cdot P(h) \cdot P(\bar{h}) \cdot P(\bar{c}\bar{q}) \cdot P(cq)\\
&\qquad + [P(e|h\bar{c}\bar{q}) \cdot P(e|\bar{h}c\bar{q}) - P(e|hc\bar{q}) \cdot P(e|\bar{h}\bar{c}\bar{q})] \cdot P(h) \cdot P(\bar{h}) \cdot P(\bar{c}\bar{q}) \cdot P(c\bar{q}))\\
&\quad = \text{sign}\left(\left(\frac{P(e|hc\bar{q})}{P(e|\bar{h}c\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right) \cdot P(e|\bar{h}c\bar{q}) \cdot P(e|\bar{h}cq) \cdot P(c\bar{q}) \cdot P(cq)\right.\\
&\qquad + \left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right) \cdot P(e|\bar{h}\bar{c}\bar{q}) \cdot P(e|\bar{h}cq) \cdot P(\bar{c}\bar{q}) \cdot P(cq)\\
&\qquad \left. + \left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hc\bar{q})}{P(e|\bar{h}c\bar{q})}\right) \cdot P(e|\bar{h}\bar{c}\bar{q}) \cdot P(e|\bar{h}c\bar{q}) \cdot P(\bar{c}\bar{q}) \cdot P(c\bar{q})\right)\\
&\quad = \text{sign}\left(\left(\frac{P(e|hc\bar{q})}{P(e|\bar{h}c\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right) \cdot \frac{P(e\bar{h}c\bar{q})}{P(\bar{h}c\bar{q})} \cdot \frac{P(e\bar{h}cq)}{P(\bar{h}cq)} \cdot P(c\bar{q}) \cdot P(cq)\right.\\
&\qquad + \left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right) \cdot \frac{P(e\bar{h}\bar{c}\bar{q})}{P(\bar{h}\bar{c}\bar{q})} \cdot \frac{P(e\bar{h}cq)}{P(\bar{h}cq)} \cdot P(\bar{c}\bar{q}) \cdot P(cq)\\
&\qquad \left. + \left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hc\bar{q})}{P(e|\bar{h}c\bar{q})}\right) \cdot \frac{P(e\bar{h}\bar{c}\bar{q})}{P(\bar{h}\bar{c}\bar{q})} \cdot \frac{P(e\bar{h}c\bar{q})}{P(\bar{h}c\bar{q})} \cdot P(\bar{c}\bar{q}) \cdot P(c\bar{q})\right)\\
&\quad = \text{sign}\left(\left(\frac{P(e|hc\bar{q})}{P(e|\bar{h}c\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right) \cdot P(e|\bar{h}c\bar{q}) \cdot P(e|\bar{h}cq) \cdot P(c\bar{q}) \cdot P(cq)\right.\\
&\qquad + \left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hcq)}{P(e|\bar{h}cq)}\right) \cdot P(e|\bar{h}\bar{c}\bar{q}) \cdot P(e|\bar{h}cq) \cdot P(\bar{c}\bar{q}) \cdot P(cq)\\
&\qquad \left. + \left(\frac{P(e|h\bar{c}\bar{q})}{P(e|\bar{h}\bar{c}\bar{q})} - \frac{P(e|hc\bar{q})}{P(e|\bar{h}c\bar{q})}\right) \cdot P(e|\bar{h}\bar{c}\bar{q}) \cdot P(e|\bar{h}c\bar{q}) \cdot P(\bar{c}\bar{q}) \cdot P(c\bar{q})\right).
\end{aligned}
$$

$$\square$$

References

[1] L. Bovens and S. Hartmann, "Bayesian Epistemology", Oxford University Press, Oxford, 2003.

[2] R. Carnap, *On the application of inductive logic*, Philosophy and Phenomenological Research **8** (1947), 133–148.

[3] L. Casini and J. Landes, *Confirmation by robustness analysis. A Bayesian account*, Erkenntnis **89** (2024), 367–409.

[4] M. Egger, G. D. Smith, M. Schneider and C. Minder, *Bias in meta-analysis detected by a simple, graphical test*, BMJ **315** (1997), 629–34.

[5] M. Fagerland, "Evidence-Based Medicine and Systematic Reviews", Research in Medical and Biological Sciences: From Planning and Preparation to Grant Application and Publication, 2015, 431–61.

[6] M. Flacco, L. Manzoli, S. Boccia, L. Capasso, K. Aleksovska, A. Rosso, G. Scaioli, C. De Vito, R. Siliquini, P. Villari and J. Ioannidis, *Head-to-head randomized trials are mostly industry sponsored and almost always favor the industry sponsor*, Journal of Clinical Epidemiology **68** (2015), 811–20.

[7] L. S. Friedman and E. D. Richter, *Relationship between conflicts of interest and research results*, Journal of General Internal Medicine **19** (2004), 51–56.

[8] J. Fuller, *Meta-research evidence for evaluating therapies*, Philos. Sci. **85** (2018), 767–80.

[9] D. Gomes and C. Stavropoulou, *The impact generated by publicly and charity-funded research in the united kingdom: a systematic literature review*, Health Research Policy and Systems **17** (2019), 22.

[10] L. Hartling, M. Ospina, Y. Liang, D. M. Dryden, N. Hooton, J. Krebs Seida and T. P. Klassen, *Risk of bias versus quality assessment of randomised controlled trials: cross sectional study*, BMJ **339** (2009), b4012.

[11] J. Higgins and S. Green eds., "Cochrane Handbook for Systematic Reviews of Interventions", The Cochrane Collaboration, 2011, version 5.1.0 edition.

[12] J. Ioannidis, *Why most published research findings are false*, PLoS Med **8** e124.

[13] J. Ioannidis, *Why most discovered true associations are inflated*, Epidemiology **19** (2008), 640–48.

[14] J. P. A. Ioannidis, D. Fanelli, D. D. Dunne and S. N. Goodman, *Meta-research: Evaluation and Improvement of Research Methods and Practices*, PLOS Biology **13** (2015), e1002264.

[15] A. W. Jørgensen, K. L. Maric, B. Tendal, A. Faurschou and P. C. Gøtzsche, *Industry-supported meta-analyses compared with meta-analyses with non-profit or no support: Differences in methodological*

quality and conclusions, BMC Medical Research Methodology **8** (2008), 60.

[16] P. Jüni, A. Witschi, R. Bloch and M. Egger, *The hazards of scoring the quality of clinical trials for meta-analysis*, JAMA **282** (1999), 1054–60.

[17] L. L. Kjaergard and B. Als-Nielsen, *Association between competing interests and authors' conclusions: epidemiological study of randomised clinical trials published in the BMJ*, BMJ **325** (2002), 249.

[18] J. Lexchin, *Those who have the gold make the evidence: How the pharmaceutical industry biases the outcomes of clinical trials of medications*, Sci Eng Ethics **18** (2012), 247–61.

[19] J. Lexchin, L. Bero, B. Djubegovic and O. Clark, *Pharmaceutical industry sponsored research: Evidence for a systematic bias*, BMJ **326** (2003), 1167–70.

[20] A. Lundh, J. Lexchin, B. Mintzes, J. B. Schroll and L. Bero, *Industry sponsorship and research outcome (Review)*, Cochrane Database of Systematic Reviews (2017), Art. No. MR000033.

[21] J. H. Montgomery, M. Byerly, T. Carmody, B. Li, D. R. Miller, F. Varghese and R. Holland, *An analysis of the effect of funding source in randomized clinical trials of second generation antipsychotics forthe treatment of schizophrenia*, Controlled Clinical Trials **25** (2004), 598–612.

[22] B. Osimani and J. Landes, *Varieties of Error and Varieties of Evidence in Scientific Inference*, British J. Philos. Sci. **74** (2023), 117–170.

[23] T. Reynolds, *Industry-funded versus publicly funded trials: Are the standards the same?*, JNCI: Journal of the National Cancer Institute **93** (2001), 1590–1592.

[24] M. Roseman, K. Milette, L. Bero, J. Coyne, J. Lexchin, E. Turner and B. Thombs, *Reporting of conflicts of interest in meta-analyses of trials of pharmacological treatments*, JAMA **305** (2011). 1008–17.

[25] D. Sackett, W. Rosenberg, J. Gray, R. Haynes and W. Richardson, *Evidence based medicine: what it is and what it isn't*, BMJ **312** (1996), 71–2.

[26] B. J. Shea, J. M. Grimshaw, G. A. Wells, M. Boers, N. Andersson, C. Hamel, A. C. Porter, P. Tugwell, D. Moher and L. M. Bouter, *Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews*, BMC Medical Research Methodology **7** (2007), 10.

[27] R. E. Slavin, *Best-evidence synthesis: An alternative to meta-analytic and traditional reviews*, Educational Researcher **15** (1986), 5–11.

[28] J. Stegenga, "Medical Nihilism", Oxford University Press, Oxford, 2018.

[29] P. E. Verde, *A bias-corrected meta-analysis model for combining, studies of different types and quality*, Biom. J. **63** (2021), 406–22.

[30] N. J. Welton, A. E. Ades, J. B. Carlin, D. G. Altman and J. A. C. Sterne, *Models for potentially biased evidence in meta-analysis using empirically based priors*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **172** (2009), 119–36.

[31] S. West, V. King, T. S. Carey, K. N. Lohr, N. McKoy, S. F. Sutton and L. Lux, "Systems to Rate the Strength of Scientific Evidence", Technical Report 47, Agency for Healthcare Research and Quality (US), Rockville, 2002. https://www.ncbi.nlm.nih.gov/books/NBK11930.

[32] J. Worrall, *What evidence in evidence-based medicine?*, Philos. Sci. **69** (2002), S316–30.

[33] V. Yank, D. Rennie and L. Bero, *Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study*, BMJ **335** (2007), 1202–05.

Lorenzo Casini
Jürgen Landes

Exploring intersubjective Attitudes Towards Conditionals

1. *Introduction*

In this paper, I lay the groundwork for the investigation of intersubjective attitudes and epistemic positions towards conditionals. The aim of this work is to understand what intersubjective attitudes two individuals can have towards conditional statements, in order to set the table for future studies and developments of a formal framework for conditionals that can account for these attitudes. The term 'attitude' is used in epistemology to denote the possible epistemic relation a person can have with a proposition, *i.e.*, believe it, disbelieve it or suspend judgement – I consider only these three main doxastic attitudes here. Hence, I follow this usage of the term 'attitude' and extend it to a social context, where more than one person is involved. Given the preparatory nature of this investigation, my focus is on the interaction between only two individuals.

By 'intersubjective attitudes' I mean the relation two individuals can have towards the same conditional. Two individuals can agree about a conditional (namely, they both believe the same conditional), disagree (they believe two conditionals with the same antecedent but contradictory consequents), or suspend their judgement on it. I here distinguish between *strong* disagreement, namely when an agent explicitly contradicts what another agent has said; and *weak* disagreement, *i.e.*, suspended judgement, where an agent does not directly contradict another agent's utterance or belief, but suspends judgement about it.

As in the literature on disagreement, I here use 'epistemic position' to characterise how much the speakers are informed about the topic considered. The possible epistemic positions are *superior*/*inferior*, where one speaker is more informed than the other, or *peers*, where the speakers are equally informed.

In this paper, I consider conditionals of all types, and treat them all alike, independently of the truth values of their antecedents and consequents, or the mood of the verbs involved – both criteria are currently used in the classification of conditionals. To try to explain the intersubjective attitudes, it is necessary to introduce and analyse the epistemic positions the speakers might have when considering conditionals.

The paper follows this structure: Section 2 introduces the account of conditionals to be used throughout this work. Section 3 explores three main intersubjective attitudes: agreement, disagreement, and weak disagreement or suspended judgement. Through the examination of an example and the application of the account of conditionals introduced in §2, I elucidate the intersubjective attitudes and the related epistemic positions that the agents involved may have. Section 4 serves as the conclusion, summarising the key findings of the analysis and suggesting potential directions for future research.

2. *Intersubjective attitudes*

The expression 'intersubjective attitudes' denotes the possible attitudes two or more individuals can have with respect to a proposition, statement or topic – let us say towards Ψ. The term is here introduced and is not widespread in epistemology. However, it is inspired by the use of the term 'attitude' in epistemology, to refer to the possible epistemic relation a *singular* individual can have towards Ψ. The individual can believe Ψ, disbelieve it, she can be certain of it, suspend judgement, etc. I transpose the term 'attitude' from standard epistemology to social contexts.

Despite the great interest in the epistemic relations that two or more agents can have towards Ψ, like agreeing or disagreeing about it, there is not, to my knowledge, a specific term to refer to these relations. However, such relations, like that of two people agreeing on Ψ, presuppose that, at least, the two agents involved *believe* Ψ. Namely, in agreeing on Ψ, the two agents have, individually, a singular attitude towards Ψ. For this reason, 'attitude' seems a legitimate term also in this context.

In social epistemology, we find expressions like 'collective knowledge', 'collective attitude', 'collective intentionality', etc. However, 'collective' is used in the literature to ascribe to a group of people doxastic or bouletic attitudes, decisions, etc. as if the group were an individual (*e.g.* see [11, 17]) – hence, no disagreement among the members of the group.[1] The term is not used to refer to the doxastic relations that two or more people can have towards a proposition, a topic, etc. with respect to each other. For this reason, 'collective' is not suitable for the purposes of the present investigation.

[1] To be more precise, there could be disagreement among members of the group, but that it not the object of study of those interested in collective attitudes. In collective attitudes, the possible disagreement has been already overcome, so the the group acts, in a broad sense, as an individual.

There is no such standardised usage for the term 'intersubjective' in the literature. It can be found in [7] in the context of confirmation theory. There, it is used to refer to the common betting quotient social group arrives at after discussion. Hence, the term is employed in a similar way as 'collective' is in social epistemology. However, unlike for 'collective', this meaning of 'intersubjective' is not settled by the scholarly practice. Moreover, it is preferable to others terms, like 'dialogical' or 'mutual'. Unlike 'dialogical', 'intersubjective' does not so strongly connote openness towards the other(s). 'Mutual', on the other hand, is used, in current epistemology, to characterise knowledge that is shared by several agents. Unlike 'common knowledge', mutual knowledge does not necessarily imply that each agent knows about other agents' knowledge (*cf.* [19] and [16]). But this seems a crucial feature of attitudes like agreement and disagreement: in order to agree (or disagree) each agent knows the other agent's belief concerning the matter under consideration. It is clear that if other people's beliefs on a given topic are not (made) explicit, it is not possible to establish whether we agree or disagree on that topic – at least, this is the standard approach in works on disagreement (for an overview, see [6]).[2] In this sense, intersubjective attitudes can be said to be 'reciprocal': agreement (or disagreement) between two agents can occur only if each of them agrees (or disagrees) with the other – it is not possible that one agrees with a person, while that person does not reciprocate.

In what follows, I focus on three main intersubjective attitudes that two individuals can have towards Ψ, where Ψ is a conditional, *i.e.* $P \rightarrow Q$. To carry out the analysis of these attitudes, I rely on current work on disagreement. In particular, I analyse intersubjective attitudes according to the epistemic positions each of the agents involved has, namely how they are positioned with respect to the proposition or topic they are discussing. There are several ways in the literature to determine someone's epistemic position. For present purposes, I focus on the background information and knowledge each of the agents has. In order to do so, I use an account of conditionals that presupposes that implicit information are involved in the assertion or acceptance of a conditional. I present this view in the next section.

3. *The account of conditionals*

Before starting the discussion and analysis of what I have called intersubjective attitudes and the related epistemic positions, it is necessary to spend

[2] For instance: 'One fairly common situation that may present opportunities for improvement is that of discovering that another person's belief on a given topic differs markedly from one's own' [2, pp. 187-8].

a few words on the (rather informal, for the moment) account of conditionals I will work with throughout the paper.

For the purposes of this paper, I will rely on an approach to conditionals inspired by [13], as presented in [18].[3] The core idea of this approach is that conditionals are

(1) enthymematic;
(2) instances of the generalisations and laws that a person believes.[4]

By (1), I mean that whenever a person utters a conditional, she is implicitly introducing some background, contextual information that she has, which she adds to the antecedent of the conditional in order to assert the consequent. Namely, in saying 'if it rains, then I'll take the umbrella' there are some implicit contextual information that are presupposed, which leads me (in conjunction with the antecedent of the conditional) to conclude that I should take the umbrella – trivially, in this case, the information might be the fact that I am going out. Of course, this type of account of conditionals is not peculiar of [13], but it was quite common at the beginning of the contemporary discussion on conditionals. For instance, [1, 8, 14] all propose theories where, in order to evaluate the truth or acceptability of conditionals, the additional implicit information are to be made explicit and added to the antecedent of the conditional under evaluation.[5] More recently, this idea has been taken up and developed into a proper semantics by [20] and [10], who employ different formal tools. However, they both agree that a conditional often has implicit propositions conjoined to the antecedent explicitly stated.[6] Following this idea, here, I formalise, loosely, a conditional as follows – I employ first-order logic to formalise conditionals and generalisations, although, for the moment, the technical details are not the primary focus. Assuming that the uttered conditional is $P \rightarrow Q$, its 'extended' version, with the implicit information Ψ made explicit, is $P \wedge R \rightarrow Q$. In this specific case, the implicit

[3] This account is developed in [15], with a probabilistic semantics, and in [9], using a possible world semantics.

[4] In this paper I do not distinguish between different types of conditionals, as many current theories do, since it is not relevant at this stage of investigation. In the literature, conditionals are classified into indicatives and subjunctives, according to the moods of the verbs involved in the sentences. Subjunctives are commonly associated with counterfactual conditionals, *i.e.* conditionals with false antecedent, as this type of conditionals is frequently expressed using the subjunctive mood.

[5] To be more precise, [1] and [8] are concerned solely with counterfactuals, and not with indicative conditionals.

[6] The types of propositions conjoined to the antecedent vary from account to account; that is, they may include laws and universal statements, as well as singular or particular propositions.

information Ψ is constituted only by R. For the purposes of this work, I assume that the set of implicit information Ψ can contain only atomic formulas or conjunctions of atomic formulas, namely $\Psi = \bigwedge_{n \geq 1} R_n$. P, R, Q, S (as well as all the following capital roman letters in this paper) stand for sentential variables. I will use lowercase letters like a, b, m, n, to denote individual constants, while I will use x, y, z, for individual variables. Furthermore, in what follows, sentential variable might be indexed with individual constants, standing for the speaker who is uttering the conditional. For instance, if the speaker is Anna and she says 'if Maria goes out, she will get a cold', the formalisation will be $O_a(m) \rightarrow C_a(m)$ (where $O(m)$ is 'Maria goes out' and $C(m)$ is 'Maria gets a cold').

(2) means that a conditional is uttered by someone because it is an instance of a generalisation that person believes. For the aims of this work, it is not necessary to make specific assumptions on generalisations – in particular, it is not necessary to strictly adhere to [13]'s account of generalisations (see [9] and [18]).[7] However, it is important to address two things. First, such generalisations have conditional form. Namely, a generalisation like 'all men are mortal' is formalised as $\forall x(H(x) \rightarrow M(x))$ – assuming that H stands for 'being a man' and M for 'being mortal'. No assumption is here made on what type of implication $\rightarrow$ is, since it is not relevant for the current discussion.[8] Of course, this applies also when the conditional is considered in its 'extended' version. In this case, the conditional is an instance of a generalisation whose antecedent is compounded by both the antecedent of the conditional *and* the implicit assumptions made by a speaker. The generalisation is formally obtained from a conditional by straightforwardly replacing the individual constant with a variable bounded by the universal quantifier.

Second, these generalisations in [13] are believed by an agent because they are deemed reliable, specifically due to their proven track record in the agent's past experience, leading to true beliefs. [13]'s generalisations also include statistical laws, namely generalisations ascribing a probability value to the consequent given the antecedent. In other words, if a law asserts that 'For all x, if $A(x)$ then $B(x)$,' then a statistical law (or *chance*,

[7] For instance, for present purposes, it is not relevant to assume that generalisations are not propositions or that they are constituted by a general enunciation and a habit of belief, as done in [13].

[8] [13] claims that such generalisations, which range over an infinite domain, are not propositions; hence, if we want to embrace his view, '$\rightarrow$' cannot be interpreted as material implication. This perspective might help in avoiding the inclusion of such generalisations into the set of implicit contextual information silently added by a speaker to the antecedent of the conditionals she utters. However, this commitment is not necessary at this stage.

in [13]'s terms) asserts that 'For all x, if $A(x)$, then there is a probability π that $B(x)$', where π is a probability value.

Now that the general account of conditionals followed here has been clarified, let us move on to the topic of the paper.

4. *Intersubjective attitudes towards conditionals*

As anticipated, I here focus on three attitudes: agreement, disagreement and suspend judgement. With respect to conditionals, these are the corresponding intersubjective attitude: first, two persons can both believe $P \rightarrow Q$, hence agree about it. Second, one person can believe $P \rightarrow Q$ and the other disbelieve it, and instead believe $P \rightarrow \neg Q$, in which case we say that they disagree.[9] Third, one person can believe $P \rightarrow Q$ and the other believe neither $P \rightarrow Q$ nor $P \rightarrow \neg Q$. This third attitude is suspended judgement. In this case, the intersubjective attitude can be seen as a form of *weak* disagreement, in the sense that the second person is not directly contradicting the other, in contrast with the second case here considered, which I therefore call *strong* disagreement.

In current epistemology, the focus is on disagreement, which is trickier to explain, more relevant in everyday life, and hence overall more interesting. Accordingly, in what follows, I primarily focus on disagreement, which is also the most interesting intersubjective attitude in relation to conditionals. Agreement tends to be less intriguing as an intersubjective attitude, particularly in the context of conditionals, as it essentially boils down to asserting, accepting, or believing the same conditional. Agreement about a conditional can be explained by appealing to the fact that the two persons involved in the discussion have the same information and believe the same generalisation, leading them to assert the same conditional. It could also be that, in real life, the two individuals hold different background assumptions. Therefore, the analysis of epistemic positions proposed for disagreement in what follows equally applies to agreement.

4.1. *Disagreement*

In everyday life, it often happens that, to express our disapproval of an agent's decision, we say 'if you do that, this [the event the agent expects] will *not* happen'. Similarly, whenever we believe that an agent should have acted differently to avoid some undesired result, we say 'if you had done

[9] It might be controversial to say that a conditional is negated by negating its consequent. However, it is quite widely accepted today (*cf.* [3]) and it is in line with the account of conditionals here adopted (*cf.* [13]).

things differently, this [whatever event occurred] would not have happened'. Likewise, in public debates, politicians often criticise the decisions made by the rival party/parties, claiming that things would have been different had alternative choices been made. Consider the debates that arose during the COVID-19 pandemic regarding the necessary containment measures to halt the spread of the contagion. Scientists too often find themselves in disagreement when contemplating counterfactual scenarios. Imagine this (not-so-) fictional exchange: 'The infection would not have stopped if we had not introduced confinement'. In response, another scientist argues: 'No! If only masks and social distancing were adopted [*i.e.*, no confinement], the infection would have ended just the same'. Of course, disagreement would still persist even if conditionals were not counterfactuals but were expressed in the indicative mood.[10]

In the current literature, the types of disagreement are classified according to the agents' respective epistemic positions. However, to delve into the different epistemic positions individuals might adopt during disagreements about conditionals, we need to begin with a toy example. The example serves two primary purposes: first, to elucidate how individuals can disagree about conditionals and subsequently attempt an explanation of this phenomenon; and second, it provides the basis to start the analysis of potential epistemic positions concerning conditionals.

The situation is as follows: Anna and Bruno are considering whether Maria should go out in the afternoon. Anna says 'if Maria goes out, she will get a cold'. Bruno disagrees, stating: 'if Maria goes out, she will not get a cold'. In the evening, Anna learns that Maria, in fact, stayed at home. During a phone conversation with Bruno that evening, she says: 'if Maria had gone out, she would have gotten a cold'. Bruno replies: 'No, if Maria had gone out, she wouldn't have gotten a cold'.[11]

Now, to grasp the *strong* disagreement regarding both indicative and subjunctive conditionals, and to analyse the possible epistemic positions involved, I rely on the account of conditionals previously introduced.

4.2. *Epistemic positions*

In the current literature on disagreement, a significant classification involves the possible epistemic positions that individuals involved might

[10] Differences related to the classification of conditionals into indicatives and counterfactuals (or subjunctives) might be relevant when considering the type of background information that a speaker has in uttering a conditional, and when choosing the best formal tool needed to model and evaluate them. The exploration of these topics is deferred to future work.

[11] This example is inspired by a scenario presented by [13], where two people disagree about the potential consequences of one of them eating a cake.

have toward the object of discussion (*cf.* [4]). There can be several components to determine the epistemic position of a person (*cf.* [5]). Here, I focus on one main factor, namely the background knowledge of a speaker, for it is the most immediate element to characterise epistemic positions, especially towards conditionals, since many theories of conditionals can account for additional implicit information an agent has (see §3).

A speaker can be in a *superior* epistemic position when she is more informed than her interlocutor about the topic. In the same situation, her interlocutor would be in an *inferior* epistemic position. Being more informed about the topic of discussion in a disagreement concerning conditionals can be seen as the fact that one of the two agents has more contextual information than the other. This additional information is often left implicit — *i.e.*, not explicitly stated — and may be considered as entering into the evaluation of a conditional as conjuncts to the antecedent that, instead, is uttered. Hence, making this information explicit can help us understand what is actually happening when two people disagree about conditionals and, perhaps, sometimes determine who is right and why.[12]

Apart form superior/inferior, the most interesting epistemic position considered in the discussion on disagreement is when two individuals are *epistemic peers*. Peers are individuals who share a similar level of background knowledge and are equally informed about the topic. They are on an equal footing in terms of their knowledge, and hence their disagreements cannot be explained by appealing to their additional information.

Let us start with the analysis of the disagreement between an agent in a superior epistemic position and an agent in an inferior epistemic position.

4.2.1. *Superior and inferior disagreement*

Consider again the example. Anna utters the conditional ‘if Maria goes out, she will get a cold’.[13] Using the formalisation previously introduced, Anna’s conditional is $O_a(m) \to C_a(m)$. In contrast, Bruno says ‘if Maria goes out, she will not get a cold’, namely $O_b(m) \to \neg C_b(m)$. Now, to explain their disagreement about whether Maria will get a cold if she goes out, an introduction and reliance on the contextual implicit information each of them might have becomes necessary. For instance, suppose that Anna knows that by going out, Maria will have fun, but she also knows that

[12] For instance, it can be employed, and often it is, to identify experts. Needless to say, this is, and will always be, a very idealised scenario with idealised agents, and I am not claiming that it is *actually* possible to discover the implicit assumptions people might make in real life.

[13] As mentioned, I am not concerned here with the differences between counterfactual and indicative conditionals; therefore, this analysis applies to both the conditionals Anna utters before and after Maria’s decision not to go out. The same applies to Bruno’s conditionals.

Maria suffers from a particular pathology, for which she has a very weak immune system. Then, the extended conditional, with the implicit information made explicit, is $O_a(m) \wedge F_a(m) \wedge W_a(m) \rightarrow C_a(m)$ – with $F(m)$ for 'Maria has fun' and $W(m)$ for 'Maria has a weak immune system'. Bruno, instead, is unaware of Maria's pathology and simply thinks that going out is fun. Then, Bruno's extended conditional is $O_b(m) \wedge F_b(m) \rightarrow \neg C_b(m)$. Given this scenario, it is evident that Anna is more informed than Bruno about Maria's overall situation. Therefore, in these circumstances, Anna is in a *superior* epistemic position with respect to Bruno concerning Maria's condition and decision. Conversely, Bruno is in an *inferior* epistemic position compared to Anna, for he has no information about Maria's health.

The superior/inferior case is perhaps not very interesting, as it can be easily explained by making explicit the implicit information assumed when evaluating what will happen or what would have happened. Nonetheless, there are interesting cases in real life, for instance whenever a person contrasts and expert concerning a specific subject matter, like science.

In the following section, I focus on a more intricate scenario: disagreements between two peers.

4.2.2. *Peer disagreement*

Two epistemic peers with respect to a topic are two agents that are in the same epistemic position with respect to that topic. As noted by [6], there are various ways in which two persons can be epistemic peers. For instance, they might both have no clue about the topic under consideration. However, this possibility is not the one I am here concerned with. Rather, I focus on rational agents who share the same contextual information.

In the example, characterising Anna and Bruno as epistemic peers means assuming that they share the same additional information about the situation. More precisely, either they both know that Maria has a specific pathology, or both are ignorant about it. Consider the former option: both Anna and Bruno know that Maria has a weak immune system. Nonetheless, they still disagree about what will happen if Maria goes out, *i.e.* $O_a(m) \wedge F_a(m) \wedge W_a(m) \rightarrow C_a(m)$ against $O_b(m) \wedge F_b(m) \wedge W_b(m) \rightarrow \neg C_b(m)$. Explaining the disagreement between the two, in this case, seems more complex.

As [12, p. 332] notes: 'evidence-sharing, in the sense needed for peer disagreement, is rare, since this sense must pinpoint disagreements caused *only* by differences in how agents evaluate evidence, not in the evidence itself.' To elucidate peer disagreement concerning conditionals, relying solely on the role of additional implicit assumptions does not suffice, as two peers must share the same information. Therefore, the disagreement must be explained by relying on some other element involved in the utterance of conditionals. Given the account of conditionals endorsed and outlined in section 3, the disagreement must be connected to the generalisations each of the agents believes and that are instantiated by the conditionals they utter. Namely, it should be that Anna believes $\forall x(O(x) \wedge F(x) \wedge W(x) \rightarrow C(x))$, while Bruno believes $\forall x(O(x) \wedge F(x) \wedge W(x) \rightarrow \neg C(x))$. Assuming, as I did, that both agents are rational and that generalisations are believed when they have proved to be reliable in an agent's experience, then Anna and Bruno must have had different experiences.[14] Their different past experiences have led them to distinct beliefs: Anna believes that if someone goes out but suffers from a pathology weakening the immune system, then (it is very probable that) this someone gets a cold. On the other hand, Bruno believes that if someone goes out and suffers from a pathology weakening the immune system, then (it is very probable that) this someone does not get a cold. Hence, peer disagreement in relation to conditionals, must depend on beliefs the agents have that are not directly involved in the disagreement—beliefs that are not strictly contextual even if implicit, but rather on previously acquired and evaluated beliefs that guide their current formation of beliefs and actions (*cf.* [18] and [9]).

So far, I have considered *strong* disagreement, where one agent contradicts another agent's belief or utterance. However, there is another form of disagreement, *weak* disagreement, where an interlocutor does not directly contradict the speaker's utterance, but instead suspends judgement on the topic under discussion. I will address this in the next section.

4.3. *Weak Disagreement or Suspended Judgement*

The last intersubjective attitude towards conditionals that I am going to consider is not properly a form of disagreement as previously considered, in the sense that the one agent says P (or $P \rightarrow Q$) and the other replies $\neg P$ (or $P \rightarrow \neg Q$). Instead, it is a kind of *weak* disagreement, where an agent neither affirms nor negates another agent's utterance but rather suspends judgement. In the context of conditionals, this means that when

[14] I acknowledge that this is sketchy, but the primary goal of the present work is to classify intersubjective attitudes and epistemic positions. A more in-depth analysis of disagreements about conditionals would involve exploring degrees of belief and the probability distributions of each agent. The probability value expressed by a generalisation should ideally represent the frequency with which an agent has, throughout their life, experienced the occurrence of an event given another. This, in turn, determines which conditionals the agent would assert and accept, and to what degree. While a more detailed analysis is not currently essential for the classification, it is left for future work. Interested readers can refer to [18] and [15] for a discussion of conditionals as instances of generalisations and related probabilistic formalisation.

one agent utters $P \rightarrow Q$, the other neither replies with $P \rightarrow Q$ nor with $P \rightarrow \neg Q$.

In the example involving Anna and Bruno, it might be the case that Bruno has no opinion about what will happen (or would have happened) if Maria goes out. Assuming this is the case, when Anna says 'if Maria goes out, she will get a cold', Bruno replies 'Oh, I don't know'. One initial explanation for Bruno's response could be a lack of contextual additional information about Maria's health situation. However, this alone is not sufficient, as observed in the case of the superior/inferior, where it led Bruno to disagree with Anna by negating the consequent of Anna's conditional. Therefore, to explain Bruno's response, it is necessary to suppose that he does not have any reason to assert either of the conditionals, *i.e.*, $O \rightarrow C$ or $O \rightarrow \neg C$. According to the view of conditionals I am following here, this can be explained by arguing that Bruno believes neither $\forall x(O(x) \rightarrow C(x))$ nor $\forall x(O(x) \rightarrow \neg C(x))$ – he has no experience of such situations. Alternatively, he may believe both consequents to be equally plausible (given the antecedent): that a person may get a cold as much as she may not, given that she goes out.[15] The same explanation holds even when Bruno has the same contextual information Anna has, *e.g.* he knows that Maria is fine but also that she has a weak immune system.

Notice that Bruno disagrees with Anna, for, trivially, he does not agree with her. However, this does not constitute disagreement in the sense this term is used in the related literature (see [6]), as Bruno does not negate what Anna believes or utters. Furthermore, it is not necessary to specify Bruno's epistemic position with respect to Anna's. Whether they are be peers or in an inferior/superior relationship, Bruno's perspective of suspended judgement does not change. However, it could impact the development of Bruno's state; for instance, if he is inferior to Anna, acquiring more information may (as may not) lead Bruno to agree with Anna.

Having delved into a comprehensive exploration and analysis of the intersubjective attitudes surrounding conditionals and the associated epistemic positions of individuals, it is time to draw conclusions.

5. *Conclusion*

In this paper, I have examined three main intersubjective attitudes that individuals can have regarding conditional sentences. The primary focus was on disagreement and the associated epistemic positions that agents

[15] As before, this can be further developed using probability and degrees of belief, by replacing 'equally plausible' with 'equally probable'.

adopt in relation to the topic under discussion, while agreement was briefly discussed. Although similar epistemic positions can emerge in instances of agreement, the analysis revolved around disagreement, due to its central role in current epistemology and relevance in everyday and scientific contexts.

Two main epistemic positions were explored: the superior/inferior dynamic and the scenario where individuals are peers. Through the analysis of an example, I proposed an explanation for how these epistemic positions, along with disagreements on conditionals, can be clarified. This involves making explicit the contextual information the individuals possess and considering the uttered conditionals as instances of generalisations believed by the speakers. In the last part of the paper, I considered another possible intersubjective attitude: weak disagreement, *i.e.*, suspended judgement. In this case, a person stays neutral – does not affirm nor negate – with respect to someone saying 'if P then Q'. As in the other attitudes, the neutral agent can be in a superior, inferior, or peer position with respect to the other agent.

The investigation presented in this paper is preliminary, and further research is needed, particularly in conducting a specific study on disagreement about conditionals. Despite receiving limited attention in the literature, conditionals play a crucial role as a primary means of expressing disagreement in both everyday life and, more importantly, in scientific contexts. This future work necessitates a dedicated analysis of counterfactuals and indicatives to reveal possible differences and similarities, such as variations in the type of the implicit information added to their respective antecedent. Additionally, incorporating degrees of belief will enhance our understanding and ability to account for disagreement, particularly among peers. Once these steps are completed, it will be feasible to formulate a theory of conditionals that accommodates disagreement among agents, ultimately providing a formal account of disagreement concerning conditional sentences.

References

[1] R. M. Chisholm, *The contrary-to-fact conditional*, Mind **55** (1946), 289–307.

[2] D. Christensen, *Epistemology of disagreement: the good news*, Philosophical Review **116** (2007), 187–217.

[3] D. Edgington, *On conditionals*, Mind **104** (1995), 235–329.

[4] A. Elga, *Reflection and disagreement*, Noûs **41** (2007), 478–502.

[5] B. Frances, "Disagreement", Polity Press, 2014.

[6] B. Frances and J. Matheson, *Disagreement*, In: “The Stanford” E. N. Zalta (ed.), Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, Winter 2019 edition, 2019.

[7] D. Gillies, *Intersubjective probability and confirmation theory*, British J. Philos. Sci. **42** (1991), 513–533.

[8] N. Goodman, *The problem of counterfactual conditionals*, Journal of Philosophy **44** (1947), 113–128.

[9] M. Günther and C. Sisti, *Ramsey’s conditionals*, Synthese **200** (2022).

[10] A. Kratzer, *Partition and revision: the semantics of counterfactuals*, J. Philos. Logic **10** (1981), 201–216.

[11] C. List, *Three kinds of collective attitudes*, Erkenntnis **79** (2014), 1601–1622.

[12] K. Munn, *Counterfactual-peer disagreement*, In: “Epistemology: Contexts, Values, Disagreement. Papers of the 34th International Ludwig Wittgenstein-Symposium in Kirchberg, 2011”, C. Jäger and W. Löffler (eds.), the Austrian Ludwig Wittgenstein Society, 2007, 329–342.

[13] F. P. Ramsey, *General propositions and causality*, In: “The Foundations of Mathematics and other Logical Essays”, F. P. Ramsey (ed.), Kegan Paul, Trench, Trübner, 1929, 237–255.

[14] N. Rescher, “Hypothetical Reasoning”, Studies in logic and the foundations of mathematics. North-Holland Publishing Company, 1964.

[15] L. Rossi and C. Sisti, “Ticket to ride: Conditionals as Instances of Generalisations”, manuscript.

[16] S. R. Schiffer, “Meaning”, Oxford University Press, Oxford, 1972.

[17] D. P. Schweikard and H. B. Schmid, *Collective intentionality*, In: “The Stanford Encyclopedia of Philosophy”, E. N. Zalta (ed.), Metaphysics Research Lab, Stanford University, Fall 2021, edition, 2021.

[18] C. Sisti, *Ramsey’s lost counterfactual*, Hist. Philos. Logic **44** (2022), 311–326.

[19] P. Vanderschraaf and G. Sillari, *Common Knowledge*, In: “The Stanford Encyclopedia of Philosophy”, E. N. Zalta and U. Nodelman (eds.), Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

[20] F. Veltman, *Prejudices, presuppositions, and the theory of counterfactuals*, In: “Amsterdam Papers of Formal Grammar, 1” , J. G. M. Stokhof (ed.), 1976, 248–281.

Caterina Sisti

Confirmation bias: a *mediated* advantage to social reasoning

From an evolutionary point of view, the confirmation bias (or *my-side* bias) poses a puzzle that still survives within the literature on social reasoning: is this bias adaptive? In this paper, I plan to bring forward a hypothesis that draws from the literature on the wisdom of crowds, collective decision-making, and noise reduction in social reasoning. The main idea is that the confirmation bias might be adaptive because it constitutes a tool of epistemic vigilance that allows reasoners to arrive independently at their arguments. This represents a social advantage in groups with certain features, where some strategies are in place to mediate individual inputs into collective results that have a higher accuracy. To advocate for this thesis I will first briefly introduce the current models on the value of confirmation bias and show the objections which have been brought forward. Furthermore, I will outline some interesting insights from the literature on the wisdom of crowds and from the methods to enhance accuracy in group reasoning and decision. Then, I will describe epistemic vigilance and how the confirmation bias could be construed as a tool serving that purpose. I will finally introduce a remark on how models on social reasoning must take into account both the role of opinion-building and rule-following, coherently with the proposal of the last section.

1. *Is the confirmation bias socially adaptive?*

The confirmation bias[1] (CB) is the tendency of reasoners to readily accept reasons which align with their accepted views and discount arguments which go against them. Traditionally (and popularly) CB has been described as an impediment to reasoning, both individual and social. This tradition draws on dual processing theory, which differentiates between

[1] Within the interactionist literature on reasoning the confirmation bias is often called "my-side bias", as can be found in [37, 46]. There are minor differences between these two concepts, but for the purposes of this paper, they are not relevant.

an automatic and fast cognitive system (System 1) and a slow and effortful one (System 2) [20]. According to this distinction, System 1 functions heuristically and is plagued by biases, which make reasoners prone to mistakes.

Recently, however, some models have adopted a strictly evolutionary point of view to argue that if CB were so obviously maladaptive it would have been negatively selected. As a consequence, the pervasiveness of CB seems to indicate that it must have some adaptive function. There are a variety of these approaches, with these four being the main ones:

(a) the argumentative theory of reasoning (ATR) [36, 37];
(b) the intention alignment model [39];
(c) the reality-matching account [41];
(d) the collectivised intellectualism hypothesis [44].

All these approaches postulate that reasoning is an inherently social endeavour, which allows for the cognitive weight of such a complex activity to be distributed between the members of a group. What distinguishes these approaches from each other is what they take to be the goal of the reasoning process.

(a) The *argumentative theory of reasoning* argues that reasoning is essentially a reputational tool. When participating in the argumentative process, each reasoner is motivated to produce new reasons and make objections to defend their reputation and, thus, keep their status within the reasoning community. Accordingly, argumentation has the goal of *persuading* argumentative adversaries. Therefore the advantages CB brings mostly concern the individuals, who use this tool to heighten their reputation and their influence within the decision-making group.

CB, within ATR, has two main functions. Firstly, and chiefly, it allows reasoners to produce arguments quickly, exploiting heuristics and inductively well-founded associations. The second function of CB would be to make the reasoner as convincing as possible: the idea being that the more reasons a reasoner produces, the more their argument is solid and persuasive. The authors seem to assume that there is a significant correlation between confidence and accuracy.

While this theory has many persuasive insights, there are a few objections that have been moved to it. The most notable is that it would seem more adaptive to be able to anticipate possible objections, rather than being subsequently surprised by them. Furthermore, the positive correlation between confidence and accuracy has been disputed [19] and persuasion seems too narrow a goal for reasoning.

(b) In [39], Norman brings forward a model that he considers to be a refinement of ATR, *i.e.* the *intention-alignment* model. Norman believes that the ATR fails to explain the collective benefits of social reasoning and, thus, he develops a *cooperative* model. In this model, each reasoner engages in social reasoning to align the intentions of the group and reach a solution to a collective problem (*e.g.*, group hunts).

Norman argues that CB would allow agents to better align their intentions. However, it is not at all clear *how* CB would be helpful in this endeavour. Moreover, if the overall idea is to create uniformity within a group through references to shared outlooks, reasoning would seem too complex a tool to be pragmatically needed.

(c) Peters, on the other hand, believes that the intellectualist assumption according to which agents reason to have accurate beliefs about the world is still salient. His approach revises this assumption by turning it around. Agents produce reasons socially to shape the world for it to align with their internal models. Social reasoning, then, would be a tool meant to scaffold the external ecological space. Peters, thus, calls his approach *reality matching*: reality is made to match the previsions and judgements of reasoners.

This operation of ecological scaffolding via reasoning in Peters' view is a product of well-publicised self-fulfilling prophecies. CB is particularly valuable in this endeavour: the more evidence is gathered to support a particular claim the more weight that claim has within the reasoning space of a group and, thus, the more plausible it is for that claim to have practical consequences.

While this model has some insights that can be related to cognitive models which are growing in popularity[2], it does not take into consideration that other agents could well exploit the existence of these prophecies to deceive the original reasoner without being noticed[3].

(d) The *collectivised intellectualism hypothesis* goes in a whole other direction. Smith and Wald believe that the benefit of reasoning is a socially expressed individual benefit, specifically in the individual's increased ability to understand and model the external world.

The authors propose that, thanks to CB, reasoners within a group build as strong a case as they possibly can, and, thus, all reasoners are exposed to a variety of diverse strong cases. The idea is that people are going to better their own understanding by hearing a variety of positions and evaluating all the evidence for and against them.

[2] For instance, it is clear that the ideas of scaffolding and of trying to predict the external environment are linked to ecological models of action, [31,47], and to predictive processing models, [7,8]. However, Peters' paper does not explore this connection.

[3] For instance, a father who publicly declares that his son does not smoke and has a CB might be blinded by his CB and not be able to see the evidence that the son smokes. This kind of unbalance makes it unlikely for this model to be adaptive as it is presented.

Two things remain unclear. First, the model does not seem to account for how adversarial CB makes the argumentative model. For instance, it is not clear that individuals would consider the strong cases that have been proposed by others. Secondly, it fails to show why a position would be made stronger by CB.

In conclusion, even though they explain how to contain it [5], none of these accounts fully explain how CB could be considered adaptive [24]. If that is the case, it still has to be shown. The take-home lesson from this analysis is that what is needed is a proposal where:

1) there is a clear collective advantage to exhibiting CB;
2) this advantage must have to do not only with the status of an individual within a community, but also with decision-making and problem-solving;
3) CB is exhibited within a balanced process where deception is discouraged rather than encouraged.

2. *Outlining an alternative: content and rules come together*

Quite frequently when engaging in argumentation reasoners put forward their opinions and then stick to their guns in defending them. If the problem at hand has no clear immediate solution, it is somewhat unlikely for someone to be able to show how they are right to others. As such, by the end of the process, people might have a more complex, subtle, and reasoned view, but they are unlikely to change their view altogether. When taking part in these discussions one often has a feeling that, in these kinds of questions, there are so many known unknowns or unknown unknowns [28] that none of the arguments presented might be entirely definitive or correct. Often enough, the truth lies in the middle. If from these discussions one had to derive pragmatic conclusions, the most efficient way would be to moderate and mediate the alternatives. At this point, the question becomes: how to mediate different positions which do not merge?

This description introduces a theme: that of the interplay between opinion-building (reasons) and rule-following (formal systems) in social reasoning. In the previous section, I briefly presented the four main alternatives and the specific objections which might defeat them. However, there is a more general critique that might be moved: all these accounts are concerned with the *content* of the reasoning process. Reasoning is always described in highly informal terms and the whole idea is that it is about *claims*, *opinions*, *beliefs*. What is not taken into account are *inferential patterns*. This is because, when considering CB, what is at stake is the accumulation of *reasons*, while formal tools are well-known to be content-independent. However, I believe both insights (content and structure) should be analysed.

This brings me to the alternative I want to explore. One way to consider the interplay between content and structure is to postulate that CB might produce a benefit in terms of *content*, which can only be useful when there is a *mediation* in terms of *structures*. The proposal is the following. CB might be useful both to the individual[4] and to the reasoning group. From the point of view of the reasoning group, it might be that CB, by epistemically isolating a reasoner, makes the reasoner somewhat *independent* from the other members of the community. This, in turn, provides a benefit for the community because it has been shown (as I will present in the following section) that mediation of the independently reached claims of a crowd is often more accurate than each individual claim. The role inferential patterns would have, in such a view, is specifically in the mediation of the independently reached claims. I will elaborate further on this role of inferences in a subsequent section.

According to this hypothesis, CB would not be helpful because it creates uniformity and a shared outlook in a group, but precisely for the opposite reason: CB is helpful because it allows different views to flourish and coexist within a group independently. This approach is similar to the collectivised intellectualism hypothesis [44] in that it highlights the advantages of having a variety of different strong cases within a community. However, it differs from this hypothesis because it focuses specifically on the benefits of this situation to a community, rather than to the individual. Furthermore, an approach of this kind would abide by the requirements I have mentioned in the previous section. Firstly, CB, in this view, is a tool for communities of decision-makers, not because they strive for homogeneity (as Norman has argued), but because of the diversity it creates. Secondly, this advantage would have nothing to do with status within a community, but with optimal collective strategies for problem-solving. Finally, since reputation and persuasion do not occupy a central role, the process can be seen as generally balanced from a game-theoretical point of view. Possibilities for imbalance, like the ones that emerge in the reality-matching account, would be eliminated by a collective and individual mediation, as I will discuss in the following sections.

3. *Crowds, experts, and fools*

To argue for this hypothesis it is necessary to immediately answer two objections. First of all, it is not entirely convincing that groups or reasoners might be more accurate than individual reasoners. Secondly, the claim

[4] I believe there are also benefits to the individual, such as the avoidance of cognitive dissonance or the ease in thriving within homogeneous communities.

that CB might render reasoners independent seems completely counterintuitive. In this section, I will tackle these two problems.

3.1. *Chasing the expert doesn't make a crowd wise*

Traditionally crowds have been considered patently irrational. In the last twenty years, however, this assumption has been challenged. The notion of *wisdom of crowds* has been highly popularised and is at the heart of the interactionist view. Mercier and Sperber make it a central notion of their argumentative theory of reasoning that groups reason better than individuals. There is extensive empirical evidence for this claim [1,23,26]. The most notable example is how the success rate in answering the Wason Selection task [49] is only 10-15% when individuals are answering, while it reaches 80% when groups are asked [32].

While these results are always accepted, the mechanisms that lead groups to be more accurate than individual reasoners are not that clear. For instance, Mercier and Sperber claim that this superiority can be explained in terms of *chasing the experts*[5]. This means that a group would be more efficient than individual reasoners because within the group there might be individuals who are better prepared and can convince others and nudge them or lead them in the right direction. An example would be how cognitive reflection tests – which mislead most reasoners at a first glance – are resolved by groups where the resolved by groups where the reasoners who get it right can immediately show (and convince) others by exposing the rule behind the puzzle [34].

This is clearly what sometimes happens with logical puzzles. In these cases, once explained, the answer becomes immediately apparent. This is not the case with questions with more uncertain answers, or very complex matters which involve the knowledge of whole corpora of information. In these instances what goes on is more likely what is described in the literature which followed Surowiecki's work on the wisdom of crowds [23,48]. In this view, crowds are accurate not because they contain experts who guide others, but because they *average* the answers of individuals. Surowiecki gives multiple examples of cases in which crowds not only outperform individual reasoners but also experts themselves. The idea is that there are so many factors to take into account that, even though experts are better equipped than average reasoners, they too will often not be able to give the best possible answer to a variety of problems. This includes not only estimation problems but also cognition, coordination, and cooperation problems. Of course, the only way this wisdom of crowds can emerge

[5] This particular phrasing is not attributable to Mercier and Sperber, but comes from [27].

is if the crowd is structured in just the right way to allow for a proper mediation. What is needed is diversity, independence of judgements [11] (*i.e.* exactly what I propose might be achieved by CB) and decentralisation.

There is a statistical reason behind the efficiency of crowds: the elimination of statistical noise. Noise, as described in [21, 22], is an unwanted variability of judgements. For instance, we would call noisy a situation where different reasoners give widely different answers to the same question. Noise, in essence, is a problem of reliability and sensitivity of instruments and techniques. The loss of accuracy in judgements of individuals is a symptom of how the tools of immediate reasoning are ill-calibrated and too sensitive to a variety of irrelevant factors (both in terms of irrelevant information and psychological influences). The image Kahneman gives is that of shots that hit a target in a quasi-random fashion[6].

The most effective way to reduce noise is by taking the mean from a set of judgements. This method, together with some strategies for decision hygiene, leads to improved accuracy in decision and reasoning processes, just by eliminating occasional, systematic and personal noise. Taking the mean between individual judgements generally outperforms even experts because even the judgement of experts is noisy (for personal idiosyncrasies or random factors)[7]. One problem with this method, however, is that it only makes sense if the judgements are reached independently. Traditionally, what is meant by independence is a state where the Condorcet Jury Theorem can be applied, namely where multiple judgements both depend on a certain state of affairs X (which is the target state the judgements are supposed to be about), but do not causally affect each other. When this kind of independence fails, *i.e.* when judgements affect each other, what obtains are informational cascades. These are phenomena where reasoners express judgements in a sequential fashion, with a few reasoners being trailblazers who influence all others. Other kinds of dependency issues exist. For instance, all reasoners may express judgements which are both influenced by the target state X and by other non-articulated common causes; the judgements could be affected by common causes but not by X; or, the judgements could be self-fulfilling prophecies [14]. An interesting example of these dependency issues is the case of biases: whole groups of reasoners could be affected by group biases they are not aware of, which would qualify as a non-articulated common cause which needs to be taken into account when aggregating the various judgements.

[6] By contrast, Kahneman describes bias as a situation where the shots hit the target all in roughly the same place, but all skewed from the centre. This is they way in which bias and noise differ.

[7] Kahneman's book is focused on examples of noise in expert decision, with mentions from the fields of law, economics and science.

A tentative conclusion to these observations would be that, provided group biases are under control, group reasoning can be extremely efficient just by aggregating and mediating a (possibly high) number of independent judgements. From an evolutionary point of view, then the logical assumption is that there should be some mechanism in place to guarantee that these judgements are really independent. The proposal, as explained, is that CB does this. Specifically, I believe that the independence guaranteed by CB might be of the first kind: CB helps create a decision-making environment where judgements are dependent on the common state of affairs X, but do not causally affect each other. I will later elaborate on how the other kinds of independence might be dealt with. To avoid the problem of informational cascades and undue influences, however, there have to be two conditions. First of all, the reasoning community has to be structured as an assembly of peers. Second, people need to be not easily taken in by the opinions of others, *i.e.* they need to be not so easily convinced. The first condition is out of the scope of the paper. Regarding the second condition, on the other hand, the question is more subtle and will be discussed in what follows.

3.2. *Not so gullible after all*

One immediate objection to this account would be that CB could easily lead to epistemic bubbles, and, thus, really does not seem to be the appropriate tool to allow reasoners to make judgements independently. This point is a consequence of the assumption that if people exhibit CB they must be gullible: easily taken in by others who want to deceive them, without having a chance of improving their epistemic situation. According to Hugo Mercier, however, people are not that gullible. First of all, it is well known that even experts exhibit CB (sometimes as much as novices), therefore the immediate association between CB and poor reasoning or gullibility is not warranted. Furthermore, in [32,33], Mercier provides extensive evidence for the claim that reasoners are not that gullible nor that easily swayed. What he argues is that it would make no evolutionary sense for reasoners to be easily deceived during reasoning processes because they would have no advantage and avoid engaging in them altogether. Moreover, it is clear that information is valuable and it makes sense for reasoners to treasure it by putting it under scrutiny[8]. This insight is what drove [9] and [13] to consider reasoning as a cheater-detection or lie-detection device. Mercier shows that, for these reasons, reasoners developed a set of

[8] This is also similar to Craig's ([10]) thesis on the value of knowledge as a normative concept to put sources of information under scrutiny.

tools for cognitive vigilance. CB is part of this set: by leading reasoners to give a default higher value to their information and be wary of the information provided by others it protects them from undue external influences.

Since being suspicious and sceptical are behaviours which carry evolutionary benefits, it makes sense that the considered mechanisms effectively produce some kind of cognitive and epistemic isolation. This conclusion is in stark contrast with what is widely and intuitively believed about epistemic bubbles and the likelihood of information cascades within groups. This contrast can be dealt with by explaining that CB is not directly responsible for the participation in epistemic bubbles, but rather for the *permanence* within these bubbles. What creates epistemic bubbles are shared core beliefs which create communities based on these beliefs. What CB might do is allow these beliefs to thrive once they are held by a reasoner (rather than nudging the reasoner into these communities in the first place). The difference is subtle but salient, as it shows that reasoners join (and stay in) epistemic bubbles not because they engage in faulty reasoning, but because they have faulty beliefs in the first place. Mercier explains this by pointing out that there are two ways to hold a belief, in a philosophical tradition that follows directly Plato's doctrine on knowledge and understanding[9]. Beliefs can be held *cognitively*[10] or *reflectively* [45]. A belief held reflectively is essentially inert, as it is insulated from cognitive consequences because the reasoner does not draw from it the possible inferences that would follow. On the other hand, to hold a belief cognitively is to actually *employ* it within the reasoning process. Since reflective beliefs are basically inconsequential, it makes sense that there would be no epistemic vigilance about them (or at least, much less than with cognitive beliefs). In other words, CB, being a mechanism of epistemic vigilance, would not come into play with beliefs held in this fashion. Mercier argues that most strange beliefs are only held reflectively, and not cognitively: people are gullible concerning beliefs that do not really matter. I want to stress that the point here is that the problem does not lie in the epistemic vigilance mechanisms (for instance CB) or in the reasoning abilities, but in the initial beliefs a reasoner might have and in their willingness to engage in group reasoning with diverse reasoners. CB seems to be beneficial, whereas the problem with gullibility and epistemic bubbles seems to be social rather than epistemic.

[9] Plato, in fact, in [42] distinguished between holding a belief (*i.e.*, possessing it) and *capturing* it (*i.e.*, using it).

[10] The terminology here is mine. [45] uses the term "intuitive" instead of "cognitive", however, I believe this to be a more obscure term because it does not highlight their features. Moreover, by changing the term I hope to avoid possible references to intuitive inferences, as described in [35].

All things considered, then, if we are to trust Mercier, the empirical evidence he brings forward seems to confirm that reasoners are highly vigilant and that CB might be a tool to keep this vigilance in place and arrive at a conclusion independently.

4. *Inferences all the way: creating an advantage*

I will now provide a toy example to illustrate what I mean by *mediation*. Imagine a community of about a hundred reasoners who are engaged in discussing the most profitable way to distribute the community's resources. There are various available projects in which to invest and some of the reasoners in this community (even if it is not clear who they are) are more knowledgeable in such matters and in the strengths and weaknesses of each project. While there is some public discussion, all reasoners in the community make up their minds independently (thanks to their CB, according to this proposal). To equally distribute responsibility within the community and to achieve a more promising result, they decide to vote on the matter. Unfortunately, however, there are some issues. First of all, not only do different reasoners favour different projects, but they also want to invest different percentages of the resources in different avenues. Furthermore, all the members of the community unknowingly ignore the base rate of certain factors, *e.g.* the average profit margin of projects of that kind. Were the members of the community to just express their judgement and vote for it, the result would be highly inefficient: they would get a conclusion that is highly incoherent and biased. They decide to remedy by doing the following: they first identify the experts in the community and weigh their opinions accordingly; then, they decide to compress the various opinions in a selected subset of positions, which are independent of each other; they test the coherence and validity of these positions; they conduct a causal analysis of their decision process and correct for existing group biases (*e.g.* the base-rate fallacy). In the end, they take a vote.

All the steps that go beyond a pure collection and aggregation of data are what I describe as mediation. This mediation is a product of deliberation: I believe this is where reasoning proper (*i.e.*, formal logic or probabilistic reasoning) comes into play. I will clarify what I mean by this first by discussing an assumption in the literature on the wisdom of crowds; secondly, by showing four examples of possible roles of formal reasoning within a process of collective reasoning; thirdly by pointing out two theoretical and social advantages of this approach.

First of all, when discussing the wisdom of crowds, it is assumed that collective reasoning and deliberation have a negative influence on the outcome of the process [29], as introducing any kind of social influence in the process would undermine independence. With this in mind, a distinction is usually enforced between the wisdom of crowds approach and collective decision-making. The first approach hinges on the elimination of all kinds of social influences and causal/statistical dependencies and is adopted in a variety of estimation tasks; the second hinges on the idea that communal sharing of reasons is essential in human cooperation, and is usually focused on wider and less defined tasks. However, it has been suggested that social influence does not necessarily eliminate the positive outcome of the wisdom of crowds, provided there is diversity and accuracy in the community of reasoners [3, 25, 30]. Moreover, it has been recently shown that there might be an advantage in developing a new approach which mixes aspects of both [2, 4, 6, 12, 15, 18, 34, 38]. This could be done either in the direction of introducing deliberation within the wisdom of crowds process (both in the sense of encouraging discussion and in the sense of socially selecting experts whose opinions are then aggregated) or by emphasizing the relevance of diversity of opinions and moderation within decision making.

Thus, it could be beneficial[11] to introduce a stage of mediation in the form of collective deliberation, be it preliminary to voting or, as is common in the literature, as a reflective stage between two voting sessions. Furthermore, what could be useful is to avoid striving for widespread consensus, and to prefer strategies where consensus via deliberation is reached in small sections of the crowd with the goal of confronting these outputs with the ones given by all the other sections. The idea, then, is not to give precise rules on when this mediation should take place or on which goal it should seek, but rather to point out how it could be generally advantageous in a variety of cases and with a variety of methods, as is being done in the growing experimental literature on the subject. What remains as a take-home message in this analysis is that, while the process of decision-making could benefit from a combination of the described strategies, there always needs to be a tool which defends reasoners and community from undue influences, in the forms of informational cascades. This is in line with my proposal to consider this function of CB.

Given the mediation of deliberation does not undermine the results of the wisdom of crowds but, rather, helps it, the next research focus should be analysing what this mediation consists of. As already mentioned, I believe this mediation should be understood in terms of the use of shared logical/probabilistic structures and inferential patterns. Focusing on struc-

[11] [19] argues that the collective decision-making approach is not more efficient than the wisdom of crowds, but this is in contrast with the literature just presented. This contrast could be a product of the different configurations of the experimental designs.

tures (both those produced by individuals and those used in collective judgements and policies) during the mediation would guarantee that coherence and validity are taken into account in integrating, aggregating, and using information.

Building on the previous toy example, I will now give examples of how working on inferential structures might be beneficial in this process.

(1) At the most abstract level, the use of shared inferential structures is what basic cooperation hinges on. While the diversity of opinions (contents) within a group of decision-makers is not necessarily problematic, the lack of a shared accepted language would be. This language could be traced in the shared conditions of inferential validity accepted within the community. To give an abstract example, it is clear that, if there was a fracture within a community between those who accept the excluded middle and those who do not, there would be some sort of basic incommunicability. The shared inferential practices, *i.e.* the shared basic structures that are accepted by a community, are what allow the use, evaluation and sharing of content in an organised way. As such, it might be that what needs to be the focus of interactionist models is the *repeated articulation of inferential structures*, rather than the reasons that are articulated via these inferential structures. Thus, without a mediation given via inferential structures, the whole possibility of applying a methodology to aggregate and evaluate independent opinions would be void.

(2) Even in cases where a community were to decide to simply apply what is dictated by the wisdom of crowds approach (*i.e.* by having everyone vote or produce a judgement independently and by aggregating these judgements) a consideration of inferential structures would still be necessary. For instance, while it is easy to imagine how such a vote could take place when the goal is an estimate or the solution to a quiz, it is not at all clear how it would go about in more complex scenarios. If every single member of the community were to express their political opinion on an issue, even supposing these opinions be rendered independent by CB, the result would clearly be inefficient. What usually goes on in such cases is that opinions are solidified into a selected number of positions whose structure is highly controlled. These few positions are what reasoners then vote on.

(3) Another way in which the use of the mediation of deliberation would be useful is in the consideration of experts. This could happen in at least three ways. First of all, it is clear that in some cases (*e.g.* when confronted with the cognitive reflection test) what is relevant is the understanding of inferential and causal structures and not the specific content associated with them. In these cases, deliberation led by experts with a focus on structures is highly beneficial, as demonstrated by [34, 38]. Secondly,

while the wisdom of crowds is considered to be more accurate than any individual judgement from an expert, it has been demonstrated that aggregating judgements from selected experts yields even more accurate results [6]. Structures come into play in the recognition of experts and in the use of their expertise. From an epistemological point of view, expertise is not only a matter of knowing the right information (*i.e.* of accepting the right content) but also of understanding it, its structure, and its modal aspect [50, 51]. If this is the case, the community has to employ, test and use inferential structures to evaluate the experts and then aggregate their judgements. Finally, there is reason to believe that confidence and its meta-representation could be beneficial (or at least have an influence) to wisdom of crowds and collective decision-making [3,12,25]. Confidence, like expertise, is a product of the use and meta-understanding of cognitive structures and inferential patterns. As such, the same considerations relevant in the case of expertise are relevant when evaluating (and modulating) confidence and weighing it in collective decision processes.

(4) Perhaps more interestingly, the idea itself of the wisdom of crowds hinges on a jury theorem and the ideal of independence, but, to apply the correct aggregation method and to understand what kind of independence is in place, communities need to solve the so-called π-problem [14]. In a nutshell, reasoners must understand if they have conditionalised on the right situation, taking into account all common causes of the target phenomenon, irrelevant factors, and other influences on their judgments. In other words: to define the opinions as independent in the first place it is not only necessary for them to be unaffected by each other (which, in my proposal, is taken care of by CB), but also for them to be conditionalised on all notable factors. This conditionalisation is external and preliminary to the aggregation of the opinions themselves. An interesting example would be the case where there is a bias that affects all reasoners homogeneously. Reasoning and formal rules would then be decisive in eliminating biases such as these [15], both from a probabilistic point of view and from a logical point of view (it is clear that the mediation of structural reflection has a role in eliminating biases, as pointed out by the literature on dual processing [40]).

Aside from the practical functions highlighted just now, at least two more general advantages are linked to the focus on structures and formal reasoning. First of all, while for individual reasoners, a certain amount of incoherence is negligible [43], what is needed in a collective reasoning setting is first and foremost a *normative* approach that guarantees fairness. This normative dimension is brought about by formal reasoning, which ensures coherence and objectivity by removing idiosyncrasies. While formal reasoning rules obviously cannot help in selecting virtuous content, they can be used to either eliminate unwanted incoherence or construct arguments in a valid and acceptable manner. Secondly, this approach aligns with the insight of interactionist theories of reasoning which aim to give a proper role to formal reasoning. In most models of reasoning, fast-and-frugal reasoning is understood as a weak and often off-base version of formal reasoning [20]: there is an assumption that reasoning ideally should always be conducted following the rules provided by the latter, while there does not seem to be a distinct status or function accorded to the former[12]. The interactionist models of reasoning, on the other hand, follow the intuition that there should be a specific function for fast reasoning, and they believe this function to be opinion-building. This has two interesting theoretical consequences. First of all, fast-and-frugal reasoning and the heuristics of intuition can be considered separately, and their merits can be understood as different from those of formal, structured reasoning [16, 17]. Secondly, and even more interestingly, in a scenario such as the one proposed here and in the other interactionist models, formal reasoning has a role that is mostly post-decisional, which is in line with a great deal of empirical evidence on dual processing reasoning.

5. *Conclusion*

What can be concluded from this analysis is the plausibility of the interplay between CB, which allows reasoners to produce independent judgements, and mediation via deliberation, which allows these judgements to be used, shared, and aggregated.

This approach is encouraging for a variety of reasons. First, it is backed by empirical evidence. Secondly, it is compatible with and improves on common interactionist models by providing a new role for CB and shared inferential patterns. In the third place, it resonates with Surowiecki's moral: "Groups generally need rules to maintain order and coherence [...] Groups benefit from members talking to and learning from each other, but too much communication, paradoxically, can actually make the group as a whole less intelligent"[13].

References

[1] A. Almaatouq, A. Noriega-Campero, A. Alotaibi, P. M. Krafft, M. Moussaid and A. Pentland, *Adaptive social networks promote the*

[12] This amounts with saying that intuition is highly flawed and formal reasoning, while sometimes difficult to achieve, is always best.

[13] [48], in "Introduction", p. XIX.

wisdom of crowds, Proc. Natl. Acad. Sci. U. S. A. **117** (2020), 11379–11386.

[2] E. Baccini, Z. Christoff, S. Hartmann and R. Verbrugge, *The wisdom of the small crowd: Myside bias and group discussion*, Journal of Artificial Societies and Social Simulation **26** (2023).

[3] B. Bahrami, K. Olsen, P. E. Latham, A. Roepstorff, G. Rees and C. D. Frith, *Optimally interacting minds*, Science **329** (2010), 1081–1085.

[4] J. Becker, D. Brackbill and D. Centola, *Network dynamics of social influence in the wisdom of crowds*, Proceedings of the national academy of sciences **114** (2017), E5070–E5076.

[5] S. Bland, *An interactionist approach to cognitive debiasing*, Episteme **19** (2022), 66–88.

[6] D. V. Budescu and E. Chen, *Identifying expertise to extract the wisdom of crowds*, Management science **61** (2015), 267–280.

[7] A. Clark, "Surfing Uncertainty", Oxford University Press, New York, NY, 2016.

[8] A. Clark, "The Experience Machine", Pantheon, New York, NY, 2023.

[9] L. Cosmides, H. C. Barrett and J. Tooby, *Adaptive specializations, social exchange, and the evolution of human intelligence*, Proc. Nat. Acad. Sci. India Sect. A **107** (2010), 9007–9014.

[10] E. Craig, "Knowledge and the State of Nature: an Essay in Conceptual Synthesis", Clarendon Press, Oxford, England, 1990.

[11] Z. Da and X. Huang, *Harnessing the wisdom of crowds*, Manage. Sci. **66** (2020), 1847–1867.

[12] G. De Polavieja and G. Madirolas, *Wisdom of the confident: Using social interactions to eliminate the bias in wisdom of the crowds*, arXiv preprint arXiv:1406.7578, 2014.

[13] J.-L. Dessalles, *Reasoning as a lie detection device*, Behavioral and Brain Sciences **34** (2011), 76–77.

[14] F. Dietrich and K. Spiekermann, *Independent opinions? on the causal foundations of belief formation and jury theorems*, Mind **122** (2013), 655–685.

[15] F. Dietrich and K. Spiekermann, *Deliberation and the wisdom of crowds*, Available at SSRN 4145405, 2022.

[16] G. Gigerenzer and W. D. Gray, *A simple heuristic successfully used by humans, animals, and machines: The story of the RAF and luftwaffe, hawks and ducks, dogs and frisbees, baseball outfielders and sidewinder missiles-oh my!*, Topics in Cognitive Science **9** (2017), 260–263.

[17] G. Gigerenzer and R. Selten, "Bounded Rationality: the Adaptive Toolbox", The MIT Press, 2011.

[18] D. G. Goldstein, R. P. McAfee and S. Suri, *The wisdom of smaller, smarter crowds*, In: "Proceedings of the Fifteenth ACM Conference on Economics and Computation", 2014, 471–488.

[19] D. Hamada, M. Nakayama and J. Saiki, *Wisdom of crowds and collective decision-making in a survival situation with complex information integration*, Cogn. Res. Princ. Implic. **5** (2020), 48.

[20] D. Kahneman, "Thinking, Fast and Slow", New York: Farrar, Straus & Giroux, 2011.

[21] D. Kahneman, A. M. Rosenfield, L. Gandhi and T. Blaser, *Noise: How to overcome the high, hidden cost of inconsistent decision making*, Harvard business review **94** (2016), 38–46.

[22] D. Kahneman, O. Sibony and C. R. Sunstein, "Noise", William Collins, London, England, 2021.

[23] T. Kameda, W. Toyokawa and R. S. Tindale, *Information aggregation and collective intelligence beyond the wisdom of crowds*, Nat. Rev. Psychol. **1** (2022), 345–357.

[24] L. Koreň, *What has been explained?*, Philos. Top. **50** (2022), 213–234.

[25] A. Koriat, *When are two heads better than one and why?*, Science **336** (2012), 360–362.

[26] T. Kugler, E. E. Kausel and M. G. Kocher, *Are groups more rational than individuals? a review of interactive decision making in groups*, Wiley Interdiscip. Rev. Cogn. Sci. **3** (2012), 471–482.

[27] R. P. Larrick and J. B. Soll, *Erratum—intuitions about combining opinions: Misappreciation of the averaging principle*, Manage. Sci. **52** (2006), 309–310.

[28] D. C. Logan, *Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry*, Journal of Experimental Botany **60** (2009), 712–714.

[29] J. Lorenz, H. Rauhut, F. Schweitzer and D. Helbing, *How social influence can undermine the wisdom of crowd effect*, Proc. Nat. Acad. Sci. India Sect. A **108** (2011), 9020–9025.

[30] P. Mavrodiev, C. J. Tessone and F. Schweitzer, *Effects of social influence on the wisdom of crowds*, arXiv preprint arXiv:1204.3463, 2012.

[31] R. Menary, *Growing Minds*, In: "Habits", Cambridge University Press, 2020, 297–319.

[32] H. Mercier, *How gullible are we? a review of the evidence from psychology and social science*, Rev. Gen. Psychol. **21** (2017), 103–122.

[33] H. Mercier, "Not Born Yesterday", Princeton University Press, Princeton, NJ, 2020.

[34] H. Mercier and N. Claidière, *Does discussion make crowds any wiser?*, Cognition **222** (2022), 104912.

[35] H. Mercier and D. Sperber, *Intuitive and reflective inferences*, In: "In two Minds: Dual Processes and Beyond", Oxford University PressOxford, 2009, 149–170.

[36] H. Mercier and D. Sperber, *Why do humans reason? Arguments for an argumentative theory*, Behavioral and Brain Sciences **34** (2011), 57–74.

[37] H. Mercier and D. Sperber, "The Enigma of Reason", Harvard University Press, 2017.

[38] J. Navajas, T. Niella, G. Garbulsky, B. Bahrami and M. Sigman, *Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds*, Nature Human Behaviour **2** (2018), 126–132.

[39] A. Norman, *Why we reason: Intention-alignment and the genesis of human rationality*, Biology and Philosophy **31** (2016), 685–704.

[40] G. Pennycook, *A framework for understanding reasoning errors: From fake news to climate change and beyond* (2022).

[41] U. Peters, *What is the function of confirmation bias?*, Erkenntnis **87** (2022), 1351–1376.

[42] Plato, "Theaetetus", Penguin classics. Penguin Classics, London, England, 1987.

[43] G. Schurz and R. Hertwig, *Cognitive success: A consequentialist account of rationality in cognition*, Top. Cogn. Sci. **11** (2019), 7–36.

[44] J. J. Smith and B. Wald, *Collectivized intellectualism*, Res Philos. **96** (2019), 199–227.

[45] D. Sperber, *Intuitive and reflective beliefs*, Mind Lang. **12** (1997), 67–83.

[46] K. E. Stanovich, R. F. West and M. E. Toplak, *Myside bias, rational thinking, and intelligence*, Current Directions in Psychological Science **22** (2013), 259–264.

[47] K. Sterelny, *Minds: extended or scaffolded?*, Phenomenology and the Cognitive Sciences **9** (2010), 465–481.

[48] J. Surowiecki, "The Wisdom of Crowds", Little, Brown & Company, New York, NY, 2004.

[49] P. C. Wason, *Reasoning about a rule*, Quarterly Journal of Experimental Psychology **20** (1968), 273–281.

[50] D. A. Wilkenfeld, *Understanding as representation manipulability*, Synthese **190** (2013), 997–1016.

[51] D. A. Wilkenfeld and C. M. Johnson, *Understanding for hire*, J. Gen. Philos. Sci. **50** (2019), 389–405.

Sofia Elisabetta Walters

Fotocomposizione: CompoMat srl, Loc. Braccone, 02040 Configni (RI), Italia

How can we reason effectively in social contexts marked by uncertainty, disagreement, and incomplete information? This volume explores foundational questions at the intersection of mathematical logic and social epistemology. It brings together contributions from philosophy, logic, computer science, and psychology, examining how individuals and groups revise beliefs, resolve conflicts, and make collective decisions. Most chapters stem from the 2023 workshop Reasoning with Imperfect Information in Social Settings, held at the Scuola Normale Superiore in Pisa. The volume shows how logical methods offer powerful, unifying tools for analyzing the dynamics of information exchange and public deliberation. In doing so, it makes a case for a rigorously formal and interdisciplinary approach to reasoning in complex social environments.