April 2022

Truth Logic and Inquiry: Arguments For The Dialectical Turn in Non-monotonic Reasoning

SANJAY MODGIL [a,1],
King's College London

Abstract. Traditionally, logic specifies the conditions under which propositions are true relative to a *given* model. Said model can be understood as a proxy for a mind-independent reality that, according to correspondence theories of truth, is what ultimately makes a proposition true. However, philosophical pragmatists argue that both correspondence theories, and more contemporary 'deflationary' accounts of truth, yield little insight into how truth functions in our cognitive and social lives. Pragmatism thus advocates an account of how propositions are 'made true' by temporally extended processes of norm-governed inquiry. On this view logic's role is in articulating these norms. In this paper I argue that argumentative, or 'dialectical', formalisations of non-monotonic logics more properly articulate this role, both in individual agent reasoning and perception, and in dialogue. In doing so I also argue that this is how we *should* conceive of the relationship between truth and logic if we are to elucidate the prescriptive role of logic in regulating our individual and collaborative epistemic practices in an increasingly AI-mediated environment.

1. Introduction

This paper advances a philosophical case for the dialectical turn in non-monotonic reasoning, suggesting it to be not merely methodologically fruitful but conceptually well-founded. In doing so, it provides a principled justification for the approach inaugurated by P.M. Dung in his seminal work [12] and subsequently developed by researchers in the fields of non-monotonic logic and structured argumentation.

The non-monotonic consequences from a belief base Δ can be characterised in terms of the dialectical evaluation of arguments and defeats defined by Δ under Dung's various semantics [12]. I argue that this dialectical turn in non-monotonic logics is warranted by pragmatist critiques of correspondence and deflationary theories of truth. Correspondence theories hold that the truth of a proposition consists in its correspondence to a mind-independent reality [10]. They thereby underpin the traditional role of logic in specifying the conditions under which propositions are made true relative to a given model-theoretic proxy for the way the world could be. Deflationary theories of truth [4] eschew appeal to the metaphysically problematic assumptions of correspondence by arguing that there is nothing substantive to the notion of truth. However, according to pragmatist accounts of truth, neither of these rival theories do justice to the role that truth plays in our individual and collaborative practices of perception, reasoning, discourse and action. Contemporary pragmatist accounts of truth are rooted in the works of the late 19th century triumvirate of American Pragmatists C. S. Peirce, William James and James

[1]Corresponding Author: Sanjay Modgil (sanjay.modgil@kcl.ac.uk).

Dewey [18]. In particular Peirce emphasised the pivotal role of inquiry in establishing beliefs whose veridical credentials (i.e., whose 'truthfulness') is cashed out in terms of their stability in the face of experiential inputs and the functional role of beliefs that is thereby enabled in successfully realising our goals [14,17,19]. On this view then, what *makes* (the propositional contents of) a belief true is not some mind-independent reality, but inquiry. Accordingly, logic should therefore be in the business of articulating the norms that govern our ongoing processes of inquiry.

In Section 2, I review critiques of correspondence, deflationary, and coherence theories of truth, by way of then offering a contemporary update on Peirce's pragmatic account of truth in Section 3; one that offers a holistic account of how the truth of our beliefs, *including those formed on the basis of perception*, is established by inferentially mediated processes of inquiry. In Section 4, I then argue that monological and dialogical formalisations of non-monotonic logics, in terms of Dung's dialectical evaluation of structured arguments, answers to the role that logic should play in articulating the norms of individual and collaborative inquiry. While this argument is grounded in the preceding conceptual analysis of how truth descriptively functions in our epistemological practices, I also suggest that this is how we *should* conceive of how truth functions if we are to address the challenges presented by the use of AI in our reasoning and decision making.

A Note on Terminology The philosophical literature distinguishes between 'truthbearers' – entities such as propositions, beliefs, sentences – that have truth values, and 'truthmakers' – the aspects of reality in virtue of which those truthbearers are true, such as facts, states of affairs etc. Propositions are understood as the contents of beliefs, thoughts and assertions etc. Moreover, a distinction is often made between primary and secondary truthbearers, so that if one's primary concern is with the truth of propositions (the primary truthbearers) then beliefs (the secondary truthbearers) are true derivatively, in virtue of the truth of their propositional content. Pragmatists on the other hand typically focus on beliefs as primary truthbearers; their propositional content being true in virtue of the truth of beliefs. Finally, empirical beliefs are distinguished; beliefs primarily justified by experience—by sensory perception, observation, or other forms of experiential input.

2. Truth and Logic: A (Very) Brief Overview

Correspondence theories of truth [10] typically cite propositions as primary truthbearers:

a proposition p is true iff p corresponds to some fact

For example, in early *logical atomist* versions of correspondence (Russell and Wittgenstein) the truth of complex propositions expressed in a predicate logic language is recursively defined, via truth functional logical connectives and quantifiers, in terms of their constituent atomic propositions. The latter are true in virtue of standing in a relation of correspondence to facts that constitute a mind-independent reality. For example, (p = 'the bottle contains water') $\wedge$ (q = 'the bottle is on the table') is true iff the atomic propositions p and q are true in virtue of corresponding to facts.

The classic critique of correspondence questions the very intelligibility of using our mental faculties to articulate the nature of the *non-mental* facts, and the correspondence relation between the facts and mental propositions. Moreover, what could the practical bearings of such a relation be? What in experience could one draw on to verify such a relation? Historically, this critique of correspondence (at least in its logical atomist guise) is closely intertwined with critiques of empiricist foundationalism, according to which

knowledge rests on foundational beliefs (sometimes identified with atomic propositions) that are epistemically warranted through direct, non-inferential modes of apprehension - such as perception or observation - and are taken to correspond to facts in virtue of a 'copying' or 'mirroring' relation between proposition and world. Secondly, correspondence struggles to articulate *how one establishes* the truth of more complex theory-laden propositions (e.g., 'climate change is anthropogenic') that are less obviously confirmed by the immediacy of everyday experience and that resist a straightforward recursive appeal to the truth of their atomic constituents.

These critiques highlight a more general pragmatic concern that while correspondence is primarily concerned with a *metaphysical* account of truth – with *propositions* that are made determinately true or false by a mind independent reality – a more pragmatic epistemological focus on *what makes a belief true* is required so as to illuminate and inform how truth does and arguably should *function* in governing our doxastic and cognitive lives[2]. Indeed, this concern also applies to now widely advocated 'deflationary' theories of truth [4] that have in part been developed as a reaction to the problematic metaphysical posits of correspondence. The essential idea is that truth is a redundant concept, in that no further difference is made by ascribing the property of truth to a proposition. Hence, asserting p is equivalent to asserting 'it is true that p' (Tp), believing p equates with believing Tp, proving p equates with proving Tp, and so on. So, for someone to have the concept of truth, is, according to deflationism, to be disposed to accept all instances of the equivalence schema: *the proposition that p, is true, iff p*. That said, deflationists typically acknowledge linguistic functions of the predicate 'true' [15]. For example, if one wants to endorse all the claims made by Einstein, rather than explicitly enumerating all claims in the form of a conjunction, one can simply say that all of Einstein's claims were true. However, while pragmatists are also inclined to reject metaphysical commitments that lack genuine explanatory value, they argue [25] that having acknowledged linguistic functions of the truth predicate, the pragmatist nettle should be fully grasped so as to comprehensively account for how the concept of truth functions. Indeed, the deflationary claim that there is nothing substantive to say about truth has more than a whiff of irony given that amongst all philosophical topics, truth has traditionally been amongst the most socio-politically and culturally relevant; never more so than now, given our contemporary 'post truth' predicament[3] and advances in LLMs.

Of particular relevance to the development and use of logic in AI, is Tarski's theory of truth. Consider Tarski's 'T schema' version of the above deflationary equivalence schema – *the sentence* $\ulcorner s \urcorner$ *is true iff* s – that applies to sentences s in an object language $\mathscr{L}$ [29]. The predicate 'true' is circumscribed for use in a meta-language $\mathscr{L}^M$ that names ($\ulcorner s \urcorner$) sentences (s) in the object language $\mathscr{L}$ (e.g.,$\ulcorner$*snow is white*$\urcorner$ names *snow is white*). Tarski then argues that a theory of truth for a language $\mathscr{L}$ must entail, for each sentence s of $\mathscr{L}$, a corresponding instantiation of the T-schema. He develops just such a theory. Elaborating on logical atomism, the truth of complex sentences *in* $\mathscr{L}$ is recursively defined, ***in*** $\mathscr{L}^{\mathbf{M}}$. However, $\mathscr{L}^M$ defines truth for atomic $\mathscr{L}$ sentences, not in terms of correspondence, but by reference to a *given* interpretation; that is, in terms of reference to ob-

[2]This general concern also applies to more contemporary theories of correspondence (e.g., [1]) that depart from earlier couplings of atomist theories of correspondence with empiricist foundationalism.

[3]It is instructive to consider Bernard William's vindication of the intrinsic valuing of truth and the value's instrumental relation even to very basic human needs [32]. Orwell's 1984 dystopia compellingly dramatises Williams' vindicatory account by imagining a dystopia that expunges the very notion of truth.

jects that belong to set-theoretic extensions of primitive predicates within a given model $\mathscr{M}$ (a formal surrogate for the way the world could be). Hence $\ulcorner$*snow is white*$\urcorner$ is true iff the object o in $\mathscr{M}$ that is referred to by the name $\ulcorner$*snow*$\urcorner$, is a member of the set of objects that satisfy the predicate $\ulcorner$*white*$\urcorner$ in $\mathscr{M}$. In essence, the correspondence relation has been dissolved into semantic relations of reference between names and objects, and satisfaction relations. But then the problem of correspondence – the articulation of the relation between the mental and the non-mental – re-emerges as the problem of how to account for these relations of reference and satisfaction. (Indeed, this problem applies to deflationism in general [11]). Moreover, note that a key motivation for Tarski was to avoid the liar paradox. Self reference in $\mathscr{L}$ is avoided by restricting use of 'true' to $\mathscr{L}^M$. This means that $\mathscr{L}$ would need to be a language, say first order logic, that could only be said to be an idealisation of a *regimented* natural language (that excludes predicates such as 'true'). Yet again, argues the pragmatist, this ignores the fact that we regularly make use of the predicate 'true' and indeed the concept of truth, in our everyday practices.

Coherence theories of truth [30] provide a partial response to these pragmatic critiques by focussing on *beliefs* as primary truthbearers. Coherence maintains that a belief is true if it coheres with our other beliefs. The 'vertical' relation of correspondence between mind and world is replaced by 'horizontal' relations of coherence amongst beliefs; relations constituted by the laws of logic, so that truth is determined by the standards of rationality. The hope is that rational inquiry yields an interlocking structure of consistent beliefs that are in reflective equilibrium. However, in eschewing appeal to some inaccessible reality, coherentism aligns with the omniprescence of interpretation: nothing counts as a reason for holding a belief, except another belief. Coherence thus often falters when accounting for the constraints that observation plays in forming beliefs, since perceptual beliefs (e.g., 'the bottle contains water') are just interpretations to be evaluated in terms of coherence with other beliefs. The concern then is that a set of beliefs can be internally coherent yet completely disconnected from reality. Another traditional objection notes that while p may cohere with some internally coherent set of beliefs S, $\neg p$ may cohere with some other coherent set S'. For the avoidance of contradiction, coherentism is thereby 'accused' of segueing to the relativist 'p is true for S' and '$\neg p$ is true for S''.

3. A Pragmatist Account of Truth

As alluded to above, a pragmatist account of truth begins by articulating how the concept of truth informs the regulation of our beliefs. The following broad brush account is rooted in the works of the late 19$^{\text{th}}$ century pragmatist Charles S. Peirce [14,17,19], while also drawing on insights from contemporary inheritors of the pragmatist tradition[4].

The American pragmatists embraced both Darwinian natural selection and the scientific method in providing an account of how beliefs are regulated so as to be adapted to the successful realisation of an agent's goals or purposes in an ever changing environment. Beliefs are characterised in functional rather than representational terms; that is, as context dependent set of defeasible dispositions to reason, perceive and act in certain ways [27]. In a given context Δ (that includes one's beliefs, goals, etc.), the belief that Tweety is a bird disposes one to defeasibly infer and thereby believe that Tweety flies $(\Delta, b \Rightarrow f)$. The belief that f disposes one to perceive that Tweety flies upon encounter-

[4]Notably, Cheryl Misak [17], Hugh Price [25] and Eric Schwitzgebel [27].

ing Tweety, to assert that f is the case given the goal of updating an interlocutor's mental state (assertions being characterised as speech *acts*), to launch Tweety from one's cupped hands (given one's goal to release Tweety from captivity) and so on. Dispositions to act and reason can be construed normatively: to assert reasons that inferentially licence f when challenged as to why f, to endorse the consequences of f, to appropriately restore coherence of one's beliefs when learning that $\neg f$. In general, dispositions 'entail' expectations (qua. 'predictions') as to one's cognitive and sensory experiences: that f cohere's with one's beliefs, that one is in receipt of sense data indicating that Tweety flies when encountering Tweety or uncupping one's hands, to experience assent from an interlocutor when asserting f. Our beliefs are thus assumed 'true' or 'fit for purpose' insofar as their predictions are confirmed by experience. When predictions are confounded by experience, i.e., when experience yields positive reasons for doubting one's beliefs, inquiry proceeds so as to better adapt them to their environment (in the sense that their entailed experiential predictions are less prone to doubt and thereby fit for purpose). For example if one sees that Tweety cannot fly, or assent is not forthcoming (an interlocutor counters with $\neg f$) the resulting doubt in f instigates inquiry. Assuming inquiry deems $\neg f$ to be sufficiently reliable (e.g., one's percept is reliable or one's interlocutor is trustworthy), the experienced incoherence of one's beliefs initiates further inquiry (e.g., by the exchange of reasons in dialogue with one's interlocutor). Establishing that Tweety is a penguin (p) and so does not fly, thereby restores coherence by updating Δ to include $\neg p$ ($\Delta, b, \neg p \nRightarrow f$). Notice that the joint inquiry instigated by the failure to experience the expected assent of an interlocutor (i.e., failure to realise the goal of updating the interlocutor's mental state), not only highlights the pragmatist's idealisation of inter-subjectivity in inquiry, but also the role of behavioural dispositions in regulating our beliefs.

Consider Maya, who returning from her daily jog enters her *dimly lit* kitchen. Her beliefs Δ_M include the belief w that the semi-transparent bottle on the kitchen table contains water (which she routinely fills before jogging). But (**M1**) upon swooping up and tipping the bottle to her lips, the predicted experiential outcome of the action–the tactile sensation of water at her lips (l)–fails to manifest. Inquiring further (**M2**), she shakes the bottle but does not experience the swish of water, thereby definitively establishing that $\neg w$ so as to revise Δ_M in favour of $\neg w$. (**M3**) Maya's attention is thereby drawn to the incoherence of her beliefs, issuing in doubt as to whether she filled the bottle or that her partner Satya is not at home. She inquires so as to restore coherence, inferring to the most likely explanation that Satya is at home (t) and most likely drank the water (she distinctly recalls running the tap to wait for cold water). (**M4**) Maya inquires further as to whether t or $\neg t$. She consults her phone messages which inform her that Satya briefly returned home while she was on her run. In **M1**, the belief that w is (given her goal to rehydrate) partly constituted by the disposition to tip the bottle to her lips. Failure to experience the predicted sensory outcome l issues in doubt or uncertainty as to whether w, thereby enjoining inquiry so as to recalibrate her beliefs. This is of course paradigmatic of the scientific method. Failure of a model's predicted outcomes to match with experimental data issues in further inquiry, thereby initiating revision of one's model so as to be both internally coherent and to align with the observed data. The veracity or 'truth' of the model (in Maya's case Δ_M) is thereby further and iteratively established, so that in the limit the predicted outcomes of observation and intervention align with the domain data. Indeed, the scientific method is exploited by **M2**'s explicit act of inquiry. To summarise, the belief that β can be characterised in terms of inferential, perceptual and behavioural

dispositions that manifest experientially, as inferential and sensory expectations. Inquiry is then instigated if experience subjects β to doubt. Inquiry begins by establishing that the recalcitrant experience is sufficiently authoritative (reliable). If so, the threatened incoherence of β and $\neg\beta$ initiates further inquiry so as to resolve the doubt and restore coherence; inquiry *broadly construed* in terms of our individual and collaborative practices of reasoning (inference, abduction revision etc.), the success/failure of actions explicitly effected to inquire (**M2**) and information seeking (**M4**). Indeed, *predictive processing* models of perception [7,8] account for how our *perceptual* beliefs are also typically subject to the same *doubt- inquiry* cycle.

Predictive processing has emerged as a unifying framework across cognitive science, neuroscience, and philosophy of mind, exerting wide influence on theories of perception, action, and cognition. A brief overview begins with the observation that moment to moment, the brain is bombarded by huge amounts of noisy sense data and is faced with the task of constructing a model of the (external and internal bodily) environment (Δ) given the sensory evidence E. This is obviously intractable if one assumes such a model as being rendered 'bottom up' on the basis of passively received sensory inputs. Rather, the brain approximates Bayesian inference, computing the posterior probability $P(\Delta|E)$ given the prior probability $P(\Delta)$ that Δ models the environment and the likelihood of the inputs E given Δ ($P(E|\Delta)$). Now, based on sustained exposure to sensory inputs, the brain learns causal probabilistic models that encode the brain's assumptions about how the states of the world (the environments $\Delta_1, \ldots, \Delta_n$ that an organism typically encounters) produce different patterns of sensory signals. These models are *generative* in that they generate predictions as to the sensory inputs that would be expected if the actual environment aligns with the hypothesised prior Δ. Informally, early coarse-grained sensory signals rapidly constrain which higher-level priors Δ are most plausible, and these priors generate predicted sensory inputs. *It is these predictions that primarily constitute one's experience of the environment !* In an ongoing process of 'prediction error minimisation', only the bottom-up sensory signals that are not predicted – the 'prediction errors' – are processed, thereby instigating Bayesian updating (qua 'revision') of the model so as to be aligned with the environment (in the sense that the discrepancies between sensory predictions and sensory inputs are resolved)[5]. Schematically, if a prediction p is subject to doubt by incoming recalcitrant sensory data ('$\neg p$'), inquiry is instigated so as to resolve the prediction error. To illustrate, upon entering the kitchen, Maya's incoming sense data engenders doubt in the sensory prediction generated by Maya's prior Δ_{Kitch} (i.e., by $w \in \Delta_{Kitch}$). However, the kitchen is dimly lit and so the brain assigns a low 'precision weighting' to the error and the predicted sensory signal entailed by w persists: she mistakenly perceives that the bottle contains water. Note now an alternative prequel to this vignette in which the kitchen is *brightly* rather than dimly lit. In this case the brain assigns a *high* precision weighting to the error (informally, the data is calibrated as being reliable). Inquiry proceeds further: her attention is sub-consciously prompted to attend more closely to the empty bottle, thereby dis-confirming w and updating Δ_{Kitch} with $\neg w$.

[5]The brain's generative model (GM) is hierarchical. Loosely speaking, abstract concepts such as bottle, chair etc, are encoded at the top of the cortical hierarchy, while each lower level encodes more fine grained information so that the lowest levels (those receiving information from our sense organs) encode information about raw sensory data. Each level generates top down predictions as to the inputs from the layer below, and predictive error minimisation occurs at each level. (Note the close parallels with GMs in machine learning.)

Peirce was strikingly prescient in anticipating this account of perceptual inference. He argued that simple perceptions do not arise from the passive reception of sensory data, but from the interplay between raw sense data and inferences drawn from the brain's best guess or prior hypothesis [33]. Peirce thus anticipated Sellar's repudiation of the 'myth of the given' [28], rejecting empiricism's reliance on perceptual beliefs as given, as self-authenticating empirical foundations. Indeed, the American pragmatists rejected wholesale the quest for indubitable foundations for knowledge (including the Cartesian foundations of direct and immediate acquaintance). The upshot is a *global* fallibilist epistemology that eschews absolute certainty in favour of an evolving inter-locking web of testable fallible inferences. Global fallibilism and the integral role of inquiry underpin Peirce's claim that what makes a belief (and thereby its propositional content) true is ultimately not grounded in facts that constitute a mind independent reality. Reversing the facts-first agenda, true beliefs (loosely speaking 'facts') are the *outcome* of inquiry, and not the indubitable bedrock upon which inquiry builds. There is no appeal to standards outside processes of inquiry. The question should not be what is a fact, but what makes for *principled* inquiry, which is to say inquiries better purposed to yield beliefs that would not be perturbed by experience (i.e., subject to doubt). We then take our beliefs to be 'certain enough', or 'true', until such time as one is presented by some positive reason for doubting them, whereupon principled inquiry realigns our beliefs so as to be better adapted to successfully realising our cognitive and behavioural goals. Cheryl Misak argues [17] that Peirce's theory of truth is thus epitomised by two conditionals:

> *"if, if inquiry were pursued as far as it could go, β would be believed, then β is true"*

is a normative claim one should countenance in light of the doubt-inquiry cycle's ongoing adaptation of beliefs to the environment. That is, if β survives inquiry broadly construed, *that's all we could ask of it*. Secondly:

> *"if β is true, then, if inquiry relevant to β were pursued as far as it could fruitfully go, β would be believed*

expresses the expectation that if β is true then one would expect of β that it survives prolonged inquiry. This is a *pragmatically significant* expectation in leading us to expect something of experience, in contrast with the correspondence expectation: *'if p is true'* then we can expect that it corresponds to the world. At this point, a further remark is in order as to the kinds of open-ended doubts that pertain not to extant beliefs, but that also instigate inquiry. Suppose that after consulting the calendar (**M4**) Maya is prompted to organise dinner out with Satya. In doubt as to which is Satya's favourite restaurant, she inquiries further, but in so doing Maya so to speak 'forms' a belief. The object of inquiry is not the propositional content of some pre-existing belief. Rather it is what Peirce characterises as (in our contemporary parlance) an 'open formula' of the form $favourite\_restaurant(Satya, X)$ where X is unbound – its reference is initially underdetermined but awaits being fixed by inquiry[6].

Note that the above account of truth does not succumb to the correspondentist's objection to coherence; namely that our perceptual beliefs are untethered to the world as

[6]Propositions with unbound variables are problematic if logic provides the conditions under which sentences are made true by some model theoretic proxy for the mind-independent world assumed by correspondence. The variable X in $favourite\_restaurant(Satya, X)$ must be in the scope of a quantifier as it is assumed from the outset that the model determinately binds X to domain objects. But then $\exists X.favourite\_restaurant(Satya, X)$ is not the object of the inquiry in question; rather inquiry is purposed to identify the referent, that is to *construct* an interpretation rather than assume an interpretation from the outset.

they they are just interpretations to be evaluated in terms of coherence with other beliefs. The pragmatist riposte is that our perceptual beliefs *are* tethered to an external reality, not in virtue of their infallible epistemic authority, but in virtue of being constrained by the extent to which their predictions align with the experiential outcomes of actions and sensory sampling. But does the concern that coherence endorses multiple coherent sets of beliefs also not apply to pragmatism? Recall that a necessary (albeit not sufficient) condition for adjudging a given S as coherent, is that the relations amongst its beliefs are normatively constituted by logical relations of consistency and inferential support and entailment. Said judgement is 'synchronic' in character, in crediting coherence at a given moment in time. However, pragmatism's commitment to a 'diachronic' explication of true beliefs established by *temporally extended processes* of inquiry, suggests that these logical relations be recast as norms (recall the aforementioned normative characterisation of dispositions to justify, endorse and restore coherence) governing the *processes* of reasoning in inquiry. Indeed, given that doubt indicates that our beliefs may be maladapted to realising our goals, the very injunction to inquire is arguably *the* process norm that characterises what truth *is*[7]. When Frump asserts 'Covid is like the common cold' (c) and Tauci counters with $\neg c$, then principled inquiry *should* ensue (i.e., to content oneself with c is true for $Frump$ and $\neg c$ is true for $Tauci$ is to violate the normative injunction to inquire further and resolve the contradiction). The normative authority arises from the fact that the successful realisation of our goals is at stake. But it is a *ceteris paribus* prescription; in everyday settings other considerations may take priority or there be no utility at stake (an interlocutor counters your claim that Acton is west of Camden, but you're too busy preparing for a meeting to engage your interlocutor). On the other hand the normative injunction operates *constitutively* at the level of perception[8].

4. The Dialectical Turn in Non-monotonic Reasoning

Logics in AI traditionally assume Tarskian model-theoretic (or possible world) semantics, according to which logic does not determine which propositions are true in the actual world, but rather specifies the conditions under which sentences are true relative to a *given* interpretation in a model/possible world. This view aligns with correspondence, inasmuch as models can be viewed as formal surrogates for possible configurations of reality. However with a pragmatist account of *inquiries* as 'truthmakers' in place, we should require of logic that it articulate the diachronic norms that govern the reasoning component of inquiry. That is to say, with the norms governing the inherently dialectical processes of inference and revision that resolve the uncertainty or conflict that arises as a result of doubt. It is, I suggest, in meeting this requirement that we furnish a philosophical vindication of Dung's dialectical formalisation of non-monotonic reasoning [12].

Given a belief base Δ, structured approaches to argumentation such as $ASPIC^{+}$[23] provide guidelines for constructing and relating arguments by defeats in a Dung framework AF_{Δ} [12]. Arguments consist of conclusions deductively and/or defeasibly supported by premises that encode other beliefs, including empirically grounded beliefs that

[7] Characterising truth in terms of this normative injunction is grounded in the work of Hugh Price [25] (An in-depth development of this normative characterisation of what truth is, can be found in this presentation.)

[8] Constitutive norms define what the activity *is*. Without them, the activity would not exist as such. E.g., the rules of chess are constitutive of chess. Remove the rules and you're no longer playing chess. Similarly, what *it is* to perceive is to engage in the doubt instigated minimisation (qua. resolution) of prediction errors.

may themselves be the outcome of perceptual inference (and are thus fallible). The arguments in an AF's extension are justified or 'accepted' in the sense that they are defended against defeats from counter-arguments, and extensions are defined under different 'semantics', reflecting different dialectical standards that must be met. For example suppose an argument A concluding that Tweety flies. Recalcitrant experience – for example sensory data confounding the expectation that Tweety flies upon encountering Tweety or uncupping one hands, or the contrary assertion of an interlocutor – furnishes premises that inferentially support a counter argument B for $\neg f$. The mutually attacking arguments issue in two extensions, one containing A, the other B. The normative injunction to inquire instigates resolution of the uncertainty (conflict). Calibration of the authority (reliability) of the recalcitrant information may definitively yield a preference for one argument or the other, else one might inquire further so as to elicit further information. For example the information that Tweety is a non-flying subclass of birds, so yielding a preference for B over A so that B defeats A and one settles on a single extension justifying the belief that $\neg f$. Of course, one might later be confronted by evidence that Tweety *does* fly given that Tweety is in fact a cartoon penguin with special powers, thereby yielding an argument C that defeats B and so defends A, restoring the belief that f. Consider **M1** in which Maya's belief that w is challenged by the absent sense data supporting the conclusion that $\neg w$, but where said challenge fails given the low precision weighting assigned to the prediction error. However, w is subsequently challenged and defeated by an argument citing the failure to experience the swish of water (**M2**). Her belief that $\neg w$ then further threatens the consistency of her beliefs and inquiry proceeds (**M3&4**) to adduce arguments that ultimately settle in favour of the belief that Satya was at home and drank the water.

In contrast to preferential *model-theoretic* semantics for non-monotonic logics [16] — $\Delta \mid\!\sim \phi$ iff the preferred models of Δ semantically entail ϕ ($\Delta \models \phi$) — Dung's extensional semantics are compatible with the notion of inquiries, rather than models, as truthmakers. Indeed, for various non-monotonic logics $\mathscr{L}$ one obtains results of the form

$$\Delta \mid\!\sim_{\mathscr{L}} \phi \text{ iff } \Delta \models_{AF_\Delta} \phi$$

where $\Delta \models_{AF_\Delta} \phi$ denotes that ϕ is the conclusion of a justified argument in an (all) extension(s). The relationship with coherentism's sets of beliefs should be obvious[9]. Except to say that the assessment of whether or not a set of *arguments* constitutes an extension depends on whether the processes of inquiry – the truthmakers – that yield these extensions are appropriately governed by the logical norms that regulate the reasoning component of inquiry. That is, whether the contained arguments are mutually consistent and satisfy the standards of dialectical acceptability, and whether their conclusions are inferentially licensed by their premises. However, this assessment applies to a given set of arguments at a given moment in time, whereas we are interested in how these logical norms are operationalised diachronically, in temporally extended processes of inquiry.

We therefore turn to argument game proof theories (e.g., [22]) that operationalise these logical norms. In these games, an agent playing the role of both proponent and opponent can construct and *move* defeating arguments constructed from Δ. Arguments in the game tree can then be labelled winning or losing according to the usual argument graph labelling rules. If one complies with protocol rules that regulate the moving of arguments, then a proponent successfully establishes that an argument A for conclusion

[9] Indeed, the consistency and closure rationality postulates [6] dictate that the conclusions of arguments in an extension meet coherence criteria of *consistency*, that a belief set includes its supporting beliefs (*sub-argument closure*) and that one believes whatever is deductively entailed by what one believes (*closure under strict rules*).

ϕ is successfully defended (winning) iff $\Delta \models_{AF_\Delta} \phi$. Now note that our pragmatist account of truth implicates a number of open research challenges for structured approaches to argumentation. For example, in practice we don't inquire indefinitely, but only insofar as we are sufficiently confident that our beliefs are 'true enough' (or true for the purposes at hand as advocated in expected utility models of decision making) while acknowledging that they may be overturned by further inquiry. This suggests a research programme focussing on integrating 'credences' that express subjective probabilities as to truth of the propositional contents of beliefs, with Bayesian inference and updating so as to holistically account for predictive processing accounts of perceptual inference. Indeed, many other open research challenges are implicated[10]. However, endorsing pragmatism's idealisation of joint inquiry, I conclude with a focus on how the logical norms governing inquiry are operationalised in communicative accounts of distributed reasoning; a focus on the relatively under-explored study of *inquiry* dialogues (although see [3]) in which agents pool their knowledge and leverage their distinct strengths in reasoning together.

The formal study of argumentation dialogues owes much to the contemporary pragmatist Robert Brandom [5] and Walton and Krabbe [31]. Intuitively, the logical norms that govern the reasoning component of inquiry can be explicated in terms of normatively binding commitments that attach to speech acts in our discursive practices (recall the normative behavioural dispositions that characterise beliefs in Section 3). When asserting that f, a participant undertakes a commitment to give reasons that inferentially license f, to endorse the consequences of f, to restore consistency by retracting assertions if they are shown to be inconsistent, to challenge the assertion that f if one believes $\neg f$, and so on. Of particular note is Prakken's work on persuasion dialogues [24] in which participants construct arguments from their respective beliefs bases so as to defeat the arguments of their interlocutor. Participants can also question asserted claims and the premises of arguments moved. For example "why ψ" questions the premise of a moved argument A. A is then labelled as losing until such time as the mover of A complies with the dialogue protocol's encoding of the aforementioned normative commitments; that is by giving reasons that inferentially license ψ (i.e., moving an argument B for ψ). B then 'attacks' the why challenge and so reinstates A as winning. These dialogues generalise the above argument games, in that whether *locutions* are winning is determined by a labelling of the dialogue graph's locutions. Then, at any point in the dialogue $\mathscr{D}$, if participants comply with the dialogue's protocol, one obtains results of the form:

$$\mathscr{D}_\Delta \mid\sim_{\mathscr{L}} \phi \text{ iff } \Delta \models_{AF_\Delta} \phi$$

That is: the locution asserting ϕ in the dialogue $\mathscr{D}$'s graph is winning iff there is a justified argument concluding ϕ in the *AF obtained from the beliefs* Δ *thus far communicated in* $\mathscr{D}$. However, persuasion dialogues that establish such results (including [13]) do not support the kinds of joint *inquiry* dialogues for which we would want that:

$$\mathscr{D}_{\Delta_{1\ldots n}} \mid\sim_{\mathscr{L}} \phi \text{ iff } \Delta_1 \cup \ldots \cup \Delta_n \models_{AF_\Delta} \phi \qquad \textbf{\textit{Inq}}$$

where $\Delta_1 \ldots \Delta_n$ are the belief bases of the human and/or AI agents $Ag_1 \ldots Ag_n$ *collaboratively* inquiring so as to establish whether ϕ follows non-monotonically from their combined beliefs. Note, the emphasis on their collaborative intention licenses an assumption that would be needed to show ***Inq***; namely that agents are honest and *exhaustive* (i.e., subject to resource constraints – recall Footnote 10 – all relevant locutions that can be

[10]For example, we would want a measure of the extent to which inquiry should proceed, sufficient to guarantee rational outcomes. Indeed, recent works address the problem of bounded rational argumentation [9].

moved are moved). However, the inherently adversarial nature of persuasion dialogues would not license such assumptions. Moreover, in persuasion dialogues agents would not be expected to jointly construct individual arguments or indeed be willing to defeat their *own* arguments should they be able to do so if further information comes to light. For example, suppose a human H and her LLM personal assistant PaLL jointly deciding on a holiday itinerary. PaLL proposes 3 itineraries I_1, I_2, I_3 in the form of arguments each of which are supported by H's requirements. H expresses her doubt as to whether I_1's destination adopts unethical labour practices (where the locution expressing the doubt attacks I_1). She queries PaLL, who either does not confirm this to be the case (in which case the attack on I_1 is retracted), or confirms this is so by referencing an online report, so enabling H to construct an argument defeating I_1. Intuitively then, protocols governing such dialogues should guide the elicitation of information that can be used to instantiate an AF (on the *rhs* of ***Inq***), and such that compliance with the protocol governs communicative interactions so that ***Inq*** is satisfied. Of course, significant additional research is needed to specify and implement protocols for scaffolding joint inquiry. For example, if one asserts that $L_1 =$ "Information about Tony Blair should not be published" and this is countered by $L_2 =$ "But he is Middle East envoy", the fact that the latter is an attack on the former can be challenged, and compliance with the protocol should enforce asserting the reason *why* L_2 challenges L_1 (effectively 'completing' the enthymeme L_2 to yield an argument whose claim is the contrary of L_1).

Finally, this paper presents a dispositional characterisation of beliefs and the predictions they entail. Recalcitrant feedback from experience instigates inquiry so as to restore coherence, as evidenced by the effective realignment of predictions with experience. The veracity or 'truth' of our beliefs is thereby further and iteratively established. Now, this mirrors the resetting of LLM parameters when LLM predictions are analogously realigned by pre and post-training human feedback. However, as we anticipate reasoning collaboratively with LLMs, leveraging their distinct capacities to our advantage, the proposed dialogical scaffolding of joint human-LLM inquiry points to the prospect of ongoing feedback and compliance with the logical norms that govern these truth-conducive inquiries. Indeed, [21,2] envisage the deployment of a dialogue manager, together with LLM use of schemes and critical questions, for scaffolding *value-aligned* inquiry dialogues. Stuart Russell's proposed solution to the value alignment problem [26] proposes that AI systems remain in a constant state of uncertainty about what the human value function is. That is to say, they remain in constant doubt and should be 'truth seeking' in the sense that they are continually subject to the normative injunction to inquire and so learn our preferences. Consider for example that PaLL queries whether H would prefer I_2 to I_3 on the grounds of being less costly. H then queries PaLL as to which itinerary has a lower carbon footprint and then argues that the differential footprint is sufficiently large to warrant a preference for I_3 to I_2 that, as it were, defeats the counter-preference (as modelled in Extended AFs [20]). In this way, the joint inquiry yields information about H's ethically salient preference that updates PaLL's preference profile for H.

Conclusions: I have proposed a pragmatist account of truth that offers philosophical support for the dialectical turn in non-monotonic reasoning. I have suggested that this is how we *should* conceive of truth, if we are to support the value-aligned enculturation of LLMs in our truth seeking practices of joint reasoning and inquiry. I also hope to have thereby pointed to some research challenges to be addressed if we are to effect dialogical scaffolding for joint human-LLM inquiry.

References

[1] D. Armstrong. *Truth and Truthmakers*. Cambridge University Press, 2004.

[2] E. Bezou-Vrakatseli, O. Cocarascu, and S. Modgil. *Towards Dialogues for Joint Human-AI Reasoning and Value Alignment*. arXiv:2405.18073v1, 2024.

[3] E. Black and A. Hunter. An inquiry dialogue system. *Autonomous Agents and Multi–agent Systems*, 19(2):173–209, 2009.

[4] A.-G. Bradley and J. A. Woodbridge. *The Deflationary Approach to Truth: A Guide*. OUP, 2025.

[5] R. Brandom. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard, 1998.

[6] M. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5–6):286–310, 2007.

[7] A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36:181–204, 2013.

[8] A. Clark. *The Experience Machine: How Our Minds Predict and Shape Reality*. Penguin, 2023.

[9] M. D'Agostino and S. Modgil. A fully rational account of structured argumentation under resource bounds. In *Proc. 29th International Joint Conference on Artificial Intelligence*, pages 1841–1847, 2020.

[10] M. David. The correspondence theory of truth. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy (Summer 2025 Edition)*. Stanford University, 2025.

[11] D. Davidson. Truth and meaning. *Synthese*, 17(3):304–323, 1967.

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

[13] X. Fan and F. Toni. A general framework for sound assumption–based argumentation dialogues. *Artificial. Intelligence.*, 216(0):20 – 54, 2014.

[14] C. Hartshorne, P. W. (1931–5), and A. B. (1958), editors. *Collected Papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press, 2008.

[15] P. Horwich. *Truth*. Oxford University Press, 1990.

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

[17] C. Misak. *Truth and the End of Inquiry: A Peircean Account of Truth*. Oxford University Press, 2004.

[18] C. Misak. *The American Pragmatists*. Oxford University Press, 2013.

[19] C. Misak. The pragmatist theory of truth. In M. Glanzberg, editor, *The Oxford Handbook of Truth*, page 283–304. Oxford University Press, 2018.

[20] S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9–10):901–934, 2009.

[21] S. Modgil. Dialogical scaffolding for human and artificial agent reasoning. In *Proc. 5th Int. Workshop on AI and Cognition*, pages 58–71, 2017.

[22] S. Modgil and M. Caminada. Chapter 6 : Proof theories and algorithms for abstract argumentation frameworks. In G. Simari and I. Rahwan, editors, *Argumentation in AI*, pages 105–129. Springer, 2009.

[23] S. Modgil and H. Prakken. Abstract rule–based argumentation. In M. G. P. Baroni, D. Gabbay and L. van der Torre, editors, *Handbook of Formal Argumentation*, pages 286–361. 2018.

[24] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15:1009–1040, 2005.

[25] H. Price. Truth as convenient friction. *The Journal of Philosophy*, 100:167–190, 2003.

[26] S. Russell. *Human Compatible: AI and the Problem of Control*. Allen Lane, 2016.

[27] E. Schwitzgebel. Dispositionalism, yay! representationalism, boo! In J. Jong and E. Schwitzgebel, editors, *The Nature of Belief*. Oxford University press, forthcoming.

[28] W. Sellars. *Empiricism and the Philosophy of Mind*. Harvard University Press, 1997.

[29] A. Tarski. Der wahrheitsbegriff in den formalisierten sprachen (papers from 1923 to 1938, 2nd edition, john corcoran (ed.), oxford university press, 1956.). *Studia Philosophica*, 1:261–405, 1933.

[30] R. Walker. The coherence theory of truth. In M. Glanzberg, editor, *The Oxford Handbook of Truth*, page 219–237. Oxford University Press, 2018.

[31] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, 1995.

[32] B. Williams. *Truth and Truthfulness – An Essay in Genealogy*. Princeton University Press, 2004.

[33] A. B. Wilson. Chapter 22: Peirce, perception, and empiricism. In C. de Waal, editor, *The Oxford Handbook of Charles S. Peirce*, page 384–406. Oxford University Press, 2024.