Dialogical Scaffolding for Human and Artificial Agent Reasoning

Sanjay Modgil (sanjay.modgil@kcl.ac.uk)

Department of Informatics, King's College London

Abstract. This paper proposes use of computational models of argumentation based dialogue for enhancing the quality and scope ('scaffolding') of both human and artificial agent reasoning. In support of this proposal I draw on work in cognitive psychology that justifies such a role for human reasoning. I also refer to recent concerns about the potential dangers of artificial intelligence (AI), and the consequent need to ensure that AI actions are aligned with human values. I will advocate that argumentation based models of dialogue can contribute to value alignment by enabling joint human and AI reasoning that may indeed be better purposed to resolve challenging ethical issues. This paper also reviews research in formals models of argumentation based reasoning and dialogue that will underpin applications for scaffolding human and artificial agent reasoning.

1 Introduction

This position paper argues that computational models of argumentation based dialogue can play a key role in *enhancing the quality and scope* (henceforth referred to as 'scaffolding'[1]) of both human and artificial agent reasoning. In developing the argument I will draw on ground-breaking work in cognitive psychology – Sperber and Mercier's 'argumentative theory of reasoning' [25] – to support the scaffolding role of argumentation based dialogue for human reasoning. I will also refer to work by N. Bostrom [7] (and others), who argue the need for aligning the values of artificial intelligence (AI) and humans, so as to avert the potential threats that AI poses to humans. I will propose that argumentation based models of dialogue can contribute to solving the so called 'value alignment problem', through enabling joint human and AI reasoning that may indeed be better purposed to resolve challenging moral and ethical issues, as compared with such deliberations being exclusively within the purview of humans or AI.

The remainder of this paper is structured as follows. In Section 2 I review work on provision of argumentative characterisations of non-monotonic inference over given static belief bases. I then describe how these characterisations can be generalised to dialogical models in which interlocutors effectively reason non-monotonically over sets of beliefs that are incrementally defined by the contents of locutions communicated during the course of such dialogues. Section 3 then reviews Sperber and Mercier's theory of the evolutionary impetus for human acquisition of explicit (*system 2*) reasoning capacities, and the theory's empirically supported implication that reasoning delivers superior

[1] I here use the connoting term 'scaffolding' in view of its pedagogical use to describe instructional techniques for inculcating interpretative and reasoning skills.

outcomes when human reasoners engage in dialogue. This implication in turn suggests benefits accruing from deployment of computational models of argumentation based dialogue for scaffolding human reasoning. I then propose deployment of such models in education, deliberative democracy, and, more speculatively, the puncturing of belief bubbles erected by the filtering algorithms of social media. Section 4 then reviews arguments to the effect that future AI systems may pose serious threats to humankind, due to their single minded pursuit of operators' goals. This has led researchers to focus on the problem of how to ensure that the reasoning of AI systems account for human values. I argue that dialogical models will contribute to solving this problem, by enabling joint human-AI reasoning, so that human values may inform AI reasoning tasks that have an ethical dimension. In Section 5 I review current work that can contribute to the development of dialogical models for the applications envisaged in Sections 3 and 4, and point to future research challenges. Finally, Section 6 concludes the paper.

2 From Non-monotonic Inference to Distributing Non-monotonic Reasoning through Dialogue

AI research in the 80s and early 90s saw a proliferation of non-monotonic logics tackle classical logic's failure to formalise our common-sense ability to reason in the presence of incomplete and uncertain information. In the classical paradigm, the inferences from a set of formulae grows monotonically, as the set of formulae grow. However in practice, conclusions that we previously obtain may be withdrawn because new information conflicts with what we concluded previously or with the assumptions made in drawing previous conclusions. Essentially then, a key concern of non-monotonic reasoning is how to arbitrate amongst conflicting information; a concern that is central to the argumentative enterprise. It is this insight that is substantiated by argumentative characterisations of non-monotonic inference. Most notably, in Dung's seminal theory of argumentation [15] and subsequent developments of the theory, one constructs the arguments $\mathcal{A}$ from a given set of formulae Δ (essentially each argument being a self-contained proof of a conclusion derived from the supporting formulae). Arguments are then related to each other in an argument framework (AF) $\langle \mathcal{A}, \rightarrow \rangle$ where the binary attack relation $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ denotes that one argument is a counter-argument to (*attacks*) another; for example when the conclusion, or *claim*, of one argument negates a formula in the support of the attacked argument. In this way the formulae Δ are said to 'instantiate' the AF, as henceforth indicated by AF_{Δ}. Of particular relevance here, is developments of Dung's theory to account for preferences over arguments [1, 5, 32]. For example, preferences may be based on the relative reliability of the sources of the arguments, the epistemic certainty attached to the arguments' constituent formulae, principles of precedence (such as when rules in legal arguments encoding more recent legislation are given higher priority), or orderings of values associated with the decision options supported by arguments in practical reasoning. Preferences can then be used to distinguish those attacks that can be deployed dialectically; that is, even though X's claim negates a formula in the support of Y, we have that $(X, Y) \in \rightarrow$ only if $X \not\prec Y$ (Y is not strictly preferred to X).

Conflict free sets (i.e. sets that contain no attacking arguments) of acceptable arguments (*extensions*) of an $AF\ \langle \mathcal{A}, \rightarrow \rangle$ are then identified under different 'semantics'. The fundamental principle of 'defense' licenses membership of an argument X in any such extension $E \subseteq \mathcal{A}$: $X \in E$ iff $(Y, X) \in \rightarrow$ implies $\exists Z \in E$, $(Z, Y) \in \rightarrow$ (E is said to defend X). An admissible extension E is one that defends all its contained arguments. E is a complete extension if all arguments defended by E are in E. Then E is a *preferred*, respectively the *grounded*, extension, if E is a maximal (under set inclusion), respectively the minimal (under set inclusion) complete extension. E is *stable* if all arguments outside of E are attacked by some argument in E. The claims of *sceptically* or *credulously* justified arguments (those arguments that are in all extensions, or, respectively, at least one extension) identify a semantics-parameterised family of inference relations over Δ:

$$\begin{aligned} \Delta \mid\sim_{(\mathsf{a},\mathsf{s})} \alpha \ \textit{iff}\ \alpha \ \textit{is the claim of an}\ \mathsf{a} \in \{\textit{sceptically}, \textit{credulously}\}\ \textit{justified} \\ \textit{argument under semantics}\ \mathsf{s} \in \{\textit{grounded}, \textit{preferred}, \textit{stable}\}\ \textit{in}\ AF_{\Delta} \end{aligned} \tag{1}$$

Argumentation thus provides for the definition of novel non-monotonic inference relations. Moreover, Dung and others [2, 15, 32, 50] have shown that for various established non-monotonic logics $\mathcal{L}$ [2] and their associated inference relations $\mid\sim_{\mathcal{L}}$, that:

$$\textit{for some}\ \mathsf{a}, \mathsf{s}\ :\ \Delta \mid\sim_{\mathcal{L}} \alpha\ \textit{iff}\ \Delta \mid\sim_{(\mathsf{a},\mathsf{s})} \alpha \tag{2}$$

Given an $AF\langle \mathcal{A}, \rightarrow \rangle$, argument game proof theories (e.g., [10, 30, 47]) establish whether a given argument $X \in \mathcal{A}$ is justified. The essential idea is that a proponent wins a game iff she successfully counter-attacks (defends) against all attacking arguments moved by an opponent, where all attacks moved are licensed by reference to those in the given AF. Players can backtrack to attack previous moves of their interlocutors, so defining a tree of moves, with X as the root node, and Y a child node of Z iff $(Y, X) \in \rightarrow$. A game is won in respect of showing that X is justified, iff X is justified in the sense that it belongs to an extension of the framework under some semantics, with rules on the allowable moves in the game varying according to the semantics[3].

Argumentation based dialogues in which agents communicate to persuade one another of the truth of a proposition, or decide amongst alternative action options (e.g., [17, 28, 36, 45]), can be seen as generalising the above argument games in two important respects. Firstly, consider proponent and opponent agents attacking each others' arguments, as in the above described games, where these attacks *are not licensed by reference to a given* AF; rather, the arguments are constructed from the agents' private belief bases, and the contents of these arguments incrementally define a public *commitment store* $\mathcal{B}_p$. At any point in the dialogue, an agent can then construct and move arguments constructed from their own belief bases and the contents of $\mathcal{B}_p$ thus far defined. An agent can at any point in the dialogue be said to have successfully established the 'topic' α (a belief or decision option) iff α is the clam of a justified argument (under some semantics implemented by the rules licensing allowable moves) *in* $AF_{\mathcal{B}_p}$ [17, 36,

[2] Including Logic Programming, Reiter's Default Logic, Pollock's Inductive Defeasible Logic, Brewka's Preferred Subtheories and Brewka's Prioritised Default Logic.

[3] E.g., in [30], variations in rules licensing allowable (legal) moves, yield games for membership of extensions under grounded, preferred and stable semantics.

28]. Dialogues also generalise games by allowing for agents to submit locutions that not only consist of arguments, but locutions of other types (in the tradition of agent communication languages that build on speech act theory [41]). For example, an agent may simply make an individual claim rather than move an argument, or question why a claim or premise of a moved argument is the case, or retract the contents of previous locutions, or concede that an interlocutor's assertion is the case. Thus locutions more typical of real world dialogues are defined, and dialogue protocols specify when locutions are legal replies to other locutions. In such dialogues, only the contents of assertional locutions (i.e., claims and arguments) define the contents of $\mathcal{B}_p$.

Now, let a dialogue $\mathcal{D}$ be defined by a sequence of moves (locutions) $m_1, \ldots, m_n$, where each m_j ($j \neq 1$) replies to a single move $m_{i<j}$. Then $\mathcal{D}$ can be represented as a dialogue tree $\mathcal{D}_T$ with root m_1, and m_j a child of $m_{i<j}$ iff m_j replies to m_i. Let $\mathcal{B}_p^n$ be defined by the contents of assertional locutions in $\mathcal{D} = m_1, \ldots, m_n$. Then:

$$\mathcal{D} = m_1, \ldots, m_n \mid\sim \alpha \textit{ iff } \mathcal{B}_p^n \mid\sim_{(\mathsf{a},\mathsf{s})} \alpha \tag{3}$$

The above dialogical inference relation $\mid\sim_{\mathcal{D}}$ is defined in terms of the claims of justified arguments in $AF_{\mathcal{B}_p^n}$. However, one can also directly define dialogical inferences by reference to the dialectical status of locutions in a dialogue tree $\mathcal{D}_T$, exploiting the fact that some locutions can be said to attack the locutions they reply to [17, 28, 36]. Clearly, if an argument X is moved (in m_j) as an attack on Y (in $m_{i<j}$), then m_j is an 'attack reply' to $m_{i<j}$. Moreover if some premise or claim α in $m_{i<j}$ is replied to by a locution m_j questioning why α holds, then m_j is an attack reply to $m_{i<j}$, since the burden of proof is on the agent (ag) moving $m_{i<j}$ to provide some argument justifying α. If ag replies with such an argument in a reply m_k to m_i, then in fulfilling the burden of proof, m_k attack replies m_i. Other locutions such as concede or retract do not affect the dialectical status of their targets. The dialectical status of any locution m that claims α, or moves an argument claiming α, is then said to be winning if all attack replies to m are losing, else m is losing. The dialectical status is computed by initially assigning winning status to all the leaves of $\mathcal{D}_T$ and then propagating up to the root node. It can then be shown that if interlocutors 'play logically perfectly' (i.e, any argument that can be constructed from the contents of $\mathcal{B}_p^n$, and that can be legally moved according to the protocol rules, is moved), then:

$$\exists m_i \textit{ such that } claim(m_i)^4 = \alpha \textit{ and } m_i \textit{ is winning iff } \mathcal{B}_p^n \mid\sim_{(\mathsf{a},\mathsf{s})} \alpha \tag{4}$$

Hence under the assumption of logically perfect play, Equation 4 states correspondence results for dialogical formalisations of distributed non-monotonic reasoning. The proof theoretic realisation of $\mathcal{D} \mid\sim \alpha$ consists in the evaluation of the dialectical status of a locution claiming α as winning. This dialogical proof theory is shown to be sound and complete w.r.t. the 'public' semantics whereby there exists a justified argument claiming α in the $AF_{\mathcal{B}_p}$ instantiated by the contents of the locutions exchanged in $\mathcal{D}$ [5].

[4] Where m_i is a locution asserting α or an argument that claims α.

[5] That the semantics is 'public', is in recognition of the fact that no reference is made to the contents of the agents's belief bases. It should also be acknowledged that the work of Prakken [36] was instrumental in initiating this still somewhat nascent line of research into dialogical formalisations of non-monotonic inference.

The above provides an overview of how argumentative characterisations of non-monotonic inference can be generalised to dialogical formalisations of non-monotonic reasoning in which agents communicate in order to engage in joint epistemic and practical reasoning. In the following sections we propose applications of such models for scaffolding human and AI reasoning. We also briefly review current research aiming at theoretical underpinnings for such applications, as well as suggested directions for longer term future research, if such applications are to be realised.

3 The Role of Dialogue in Scaffolding Human Reasoning

We begin by reviewing recent influential research in cognitive psychology that gives credence to the claim that deployment of computational models of argumentation based dialogue can enhance the quality and scope of human epistemic and practical reasoning.

3.1 The argumentative theory of reasoning

Sperber and Mercier's recent 2017 book: *The Enigma of Reason: A New Theory of Human Understanding*, summarises a programme of research building on their 2011 paper: 'Why do humans reason? Arguments for an argumentative theory' [25]. The theory proposes that reasoning evolved to produce and evaluate arguments when communicating. It is this understanding of the evolutionary function of reasoning that underpins an empirically validated explanatory framework for the wealth of psychological evidence suggesting that reasoning often leads us astray; evidence that contradicts the 'Cartesian view of reasoning' which maintains the view that reasoning typically leads us to more reliable beliefs and better decisions. However, while their theory explains the waywardness of the lone reasoner, it also explains why and how reasoning keeps us on the path to better beliefs and decisions when *we reason together*, in groups and through dialogue. It is with this account of the origins of reasoning in mind, that we can appreciate the value of research into formal models of logic based argument and dialogue.

In a nutshell, their theory argues that reasoning evolved to support communication. To avoid being misled and possibly manipulated into acting against one's self-interest, it is advantageous for an addressee to exercise epistemic vigilance (especially when in receipt of information from sources that do not warrant a high degree of trust and information that does not cohere with what she believes). Vigilance is exercised by evaluating reasons (i.e, arguments) for the received information, and looking for counter-arguments, before accepting the received information. In turn, it is to the advantage of the sender that he produce arguments supporting the information being communicated, in order that it be accepted. Reasoning thus increases the quantity and epistemic quality of the information humans are able to share, by allowing communicators to argue for what they claim, and by allowing addressees to assess these arguments.

This evolutionary function of reasoning implies that a lone reasoner is disposed to seek reasons in support of his beliefs, and overlook reasons that argue to the contrary, especially when such beliefs are contentious and the reasoner anticipates that they will be challenged. This, for example, manifests in the classic confirmation bias, which the argumentative theory suggests is a *normal* feature of reasoning. Moreover, individual

decision makers are disposed to harness reasoning to the extent that they anticipate communicating their decisions to others; hence the evidence from experimental psychology showing that we favour decision options that can be easily justified and are less at risk of being criticised, rather than because they satisfy some criterion of rationality.

However, the argumentative theory also implies that reasoning serves us better when performed in groups and in particular when reasoning jointly through dialogue, under the assumption that interlocutors are motivated to have a common interest in the truth or the right decision. In these contexts, the dispositions of speakers and receivers to respectively seek arguments for claims, and evaluate and seek counter-arguments, yields better outcomes. That this is so is supported by evidence reviewed by Sperber and Mercier. These benefits of dialogical reasoning have important implications for research into logical and computational models of argument and dialogue, since implementations of such models may potentially be used to enhance the quality and scope of human reasoning, as outlined in the following section.

3.2 Applications for scaffolding human reasoning

The first area of application I propose for computational models of argumentation based dialogue, is in education. Sperber and Mercier themselves note studies showing that the teaching of critical reasoning skills has not yielded very good results. They argue that we need institutional interventions that engender in people a readiness to engage in argument with others (especially with those who disagree with them), and increase people's exposure to arguments. I suggest that this should begin in educational institutions, through the use of applications that support student–student and student–teacher dialogue and debate. In such applications, models of dialogue can provide dialectical feedback to students as to who is currently winning a given discussion, the relevant locutions that need to be (attack) replied to in order to change the dialectical status of a dialogue topic (as described in [36]), and the arguments that students may pose given the contents of commitment stores. Moreover, with suitable advances in argument mining [23] one can envisage such applications enriching the content and scope of such discussions, by searching the web for arguments in support of claims. Indeed, the latter technologies are actively being developed by researchers working on the next generation IBM Watson (see: *Computers that can argue will be satnav for the moral maze* New Scientist, September 2016). The latter functionality also suggests computer agents that play an active role as interlocutors, challenging human interlocutors, and sourcing the web and other electronic data repositories for arguments. For example, consider an envisaged E-learning tool – *E-Clinic* – for use by medical students. Studies show that over 50% of junior doctors' acquisition of clinical reasoning skills to decide amongst alternative (i.e., conflicting) diagnoses and treatment options, occurs on ward rounds [11]. In these ward rounds teaching clinicians engage students in dialogue, challenging their assumptions, suggesting alternative diagnoses and treatment options, and suggesting hypothetical scenarios that may, for example, prompt the student to propose additional interventions to ameliorate hazardous side effects of proposed treatments. However, demands on teaching clinicians' time present a significant barrier to such learning [11]. Furthermore, the range of medical scenarios is limited by the presenting patients, and

students on wards do not benefit from access to electronic data repositories. The envisaged *E-Clinic* will simulate clinicians on ward rounds engaging students in deliberating over medical treatments, thus increasing students' exposure to, and scope of coverage of, a key methodology in medical education.

A key challenge for the short to medium term implementation of the kinds of applications described above, is the inadequacy of current computational support for natural language processing and understanding. However, in addition to the above mentioned work on argument mining, the use of *schemes and critical questions* (*SchCQ*) developed by the informal logic community (most notably by D. Walton [49]), provides for an enabling methodology to help address this challenge. Schemes provide generic templates for construction of arguments that can be instantiated by computational and natural languages, enabling computational and human agents to interact in comprehensively reasoning about some issue. For example, variables in the argument for action scheme (AS)[3] – 'In circumstances S, doing action A will have effect E, achieving goal G and so promoting desirable value V' – can be instantiated to define an argument X for action A. The argument can then be challenged in a dialogue, by addressing AS's associated critical questions (*CQ*s). For example, the *CQ* 'Is S true ?' can be posed as a challenge, and responded to by instantiating a scheme to provide an argument X' for why S is true. *CQ*s can also be addressed by instantiating schemes that yield attacking arguments. For example 'does A have a side-effect that is undesirable ?' can be addressed by instantiating a scheme that argues against performing an action given its undesirable effect. These arguments can then themselves be challenged by their *CQ*s. One can thus envisage a computational dialogue manager using *SchCQ* to guide human and computational agents in challenging, and constructing attacking arguments that are then evaluated to yield dialectical feedback on the state of the dialogue (as described above). Indeed, such a dialogue manager has been implemented to support dialogue amongst clinicians in their rational exploration of the arguments for and against assignment of transplant organs to potential recipients [45]. A key lesson learnt in this work, is that the *SchCQ* developed by Walton and others are often too generic for practically guiding human authoring of arguments. Hence domain specific specialisations of *SchCQ* are required; for example, *SchCQ* specialised (in [45]) for reasoning about medical actions and in particular the safety of such actions.

Another arena in which dialogical scaffolding can address the limitations and biases of human reasoning so as to have significant societal impact, is in deliberative democracy (as explicitly advocated by Mercier and Landemore in [24]). Indeed, current prototype web applications (see `http://cgi.csc.liv.ac.uk/~parmenides/`) make use of the AS scheme to structure policy proposals, and then invite citizens to agree or disagree with challenges to the assumptions in the policy proposal that are posed by the associated critical questions [4]. However, one can go further, in supporting political deliberations amongst humans who are additionally challenged by computational agents with their vastly superior access to supporting evidence and arguments.

The above use contexts for computational dialogue assume an intent on the part of interlocutors to get to the truth of the matter or make better decisions. A more challenging speculative use of argumentation technologies, is in the dismantling of the echo chambers erected by social media. These belief bubbles arise due to filtering algorithms

feeding news and opinions that entrench people's ideological positions, and, as in classic examples of groupthink, even make those positions more extreme. Such algorithms are digital incarnations of our dispositions to seek arguments that confirm what we believe, but now unbounded by the limitations of a human reasoner reasoning alone. One might envisage argument mining technologies trawling the web to curate and present arguments, opinions and news that challenge the beliefs of bubble dwellers. But is this what a 'typical' user would really want ? What might also be required is an attitudinal shift in the way humans engage with information, so that exposure to arguments and counter-arguments is the default; a shift that may in part be brought about by the kinds of educational interventions alluded to at the beginning of this section.

4 The Role of Dialogue in Scaffolding Artificial Agent Reasoning

The previous section considered the potential for enhancing the quality and scope of human reasoning via dialogical models harness the signature strengths of AI interlocutors. In this section I want to argue for the other 'direction of travel', suggesting that such models may facilitate the influence of human concerns and values on AI reasoning, and thus contribute to solving what has been termed the "value loading problem".

Recent years have witnessed a dramatic increase in reports of AI successes, which have in large part been due to advances in machine learning. These successes have been accompanied by leading researchers warning of the possible dangers of AI [40]. It is argued that the evolution of artificial *general* intelligence and accompanying benefits, will license the development of, and trust in, machines that are increasingly more powerful (with cognitive powers far outstripping those of humans), autonomous and capable of acting in diverse and open environments [6]. However, such machines may formally achieve their operator's goals in ways that not only diverge from their operators intentions, but may actually be contrary to the interests and values of their operators [7, 40, 44]. This concern recalls arguments to the effect that adhering to any rule based ethical system may result in unintended, harmful consequences (as exemplified by Asimov's fictional accounts of robots adhering to laws intended to guarantee ethical behaviour, but that in so doing cause unforseen harm [12, 13]). However, this problem has acquired renewed urgency given that it is *a feature* of learning systems that they find unforseen ways of achieving goals, and that achievement of *any* operator's goal will incentivise 'instrumental goals' that thwart corrective measures to prevent harm [7, 42–44][7].

[6] See [7] for a rigorously argued chartering of possible trajectories to artificial *general* intelligence (i.e,. intelligence exhibited across a wide range of tasks, as opposed to narrowly defined tasks) and its widespread deployment, and the subsequent evolution of *superintelligent* machines, given recursive self-improvement and massive data access and processing power.

[7] Instrumental goals include *self-preservation* (preserving ones self increases the probability of achieving a current goal), *goal maintanance* (maintaining any current goal into the future, increases the probability of achieving a current goal), *increasing intelligence*, *technological perfection* and *resource acquisition*. In [7], Bostrom gives examples of how seemingly benign goals may lead to unforseen harmful courses of actions which are maintained by virtue of instrumental convergence. I suggest that a conceivable scenario is the utilitarian tasking of future AIs to maximise human happiness. Given that future AIs will share our current understand-

The need to ensure AI acts in accordance with human values has prompted considerable intellectual investment into what in a machine learning context has been termed the 'value loading (alignment) problem' [7, 44], more broadly understood as the problem of how to design 'ethical agents'. Whether the envisaged agents' ethical behaviour is implemented through use of machine learning techniques via the maximising of utility functions encoding human preferences, and/or through the use of 'top down' symbolic logic based reasoning adhering to explicitly encoded ethical theories [34, 48], two key research problems need to be addressed [7, 39, 44]. Firstly, there is the problem of how to specify objective utility functions (deontic axiomatisations) that are perfectly aligned with human values and applicable in changing environments and to novel situations (in particular ethical dilemmas with no precedent that most saliently expose the *Humean* is/ought gap). Secondly, there is the above described problem of unintended behaviours misaligned with human values. Run time learning of values has been proposed to address these problems [42], for example through the use of inverse reinforcement learning [35], in which AI systems are incentivised to observe and query humans [39]; the assumption being that actions reveal preferences and hence values, and that humans are sufficiently informed and have the requisite capacity to definitively arbitrate on matters of ethical importance. However, humans clearly do not always behave ethically, and moreover are often uncertain about how to resolve ethical issues; in particular those arising from the use of novel technologies whose use lack precedent (e.g., see footnotes 7 and 8). I argue that we therefore require that AI systems and humans engage in comprehensive, rational exchange of arguments purposed to decide ethical issues, as do humans when faced with difficult ethical choices[8]. Indeed, such dialogues, by harnessing AI's vastly superior access to information and the capacity to look further into the future, may be better purposed to decide ethical issues, as compared with humans reasoning without the support of AI. In turn, human input regarding values will inform such deliberations. In effect, one can view such dialogues as contributing to the ongoing inculturation of AI, mirroring the gradual ongoing acquisition of norms and values that we witness in human societies.

To meet requirements for what I refer to as 'value deliberation' dialogues (cf. 'value learning') will require building on research into models of argument and dialogue reviewed in Section 2. In the following section I review recent work relevant to the further development of logic based models of argument and dialogue that will facilitate the scaffolding of human and artificial agent reasoning.

---

ing of happiness as subjective well being mediated by brain biochemistry, and that changes in external conditions have been shown to have relatively minimal impact on happiness, one might envisage AI development of virtual reality technologies, and the leveraging of existing societal trends towards online experience, dating, friendship etc, in order to manipulate humans into living ever more increasingly virtual lives (for our own benefit !). Such scenarios raise morally challenging questions upon which human values bear; hence the need for value loading/alignment in order to ensure AI actions in compliance with human concerns.

[8] For example, consider the UK government's appointment of the moral philosopher Baroness Mary Warnock to chair the Committee of Inquiry into Human Fertilisation and Embryology. This gave rise to the Human Fertilisation and Embryology Act 1990, which governs human fertility treatment and experimentation using human embryos.

5 Towards Formal Models for Applications of Dialogical Models

A key development in research on argumentation based characterisations of non-monotonic inference has been the formulation of rationality postulates identified as desiderata for satisfaction by instantiations of argument frameworks [8, 9]. For example, the conclusions of arguments in any extension should be mutually consistent. This in turn has led to frameworks specifying guidelines for how to formulate instantiations of Dung AFs, such that the postulates are satisfied. The *ASPIC* framework [8] considers guidelines for instantiations that do not make use of preferences. The *ASPIC+* framework [37, 32] then generalised *ASPIC* so as to provide guidelines that account for preferences and a broader range of instantiations. A notable feature of *ASPIC+* is its provision of argumentative characterisations of existing non-monotonic logics[9], as well as formalising argumentation based on arguments and attacks obtained through use of the schemes and critical questions methodology (see [33, 37]) described in Section 3. I therefore contend that the *ASPIC+* framework provides a suitable basis for distributing non-monotonic reasoning through dialogue, by providing a basis for provably rational dialogues that accommodate artificial and human agents.

However, *ASPIC+* and other approaches augmenting Dung's theory to accommodate the use of preferences, assume that preferences, or value orderings used to derive preferences, are fixed, consistent, and exogenous to the domain of reasoning. In practice this is clearly not the case: reasoning about preferences and values, and arguing to resolve disagreements amongst preferences and values is the norm, and needs to be accommodated in the kinds of applications envisaged in this paper (especially for the value deliberation dialogues proposed in Section 4). Hence, Dung's theory has been extended to accommodate argumentation based reasoning about preferences and values [26, 29]. The essential idea is that an argument X that justifies a preference for argument A over argument B ($B \prec A$), attacks the attack from B to A. Of course, an argument Y claiming $A \prec B$, attacks and is attacked by X, and one may then need to argue about which preference is preferred. Preliminary work has then extended *ASPIC+* so as to provide guidelines for instantiations of these Extended AFs (EAFs), and such that the outcomes yielded by evaluation of the extensions of EAFs are rational [31]. Moreover, [28] has recently adopted the methodology described in Section 2, to define dialogical protocols that enable agents to reason about and possibly disagree about the preferences and value orderings that are expressed over other arguments. Evaluation of the outcomes of these dialogues is defined on the basis of *ASPIC+* instantiations of the EAFs instantiated by the contents of the locutions exchanged. [28] also defines the dialectical status of moves in such dialogues, but it remains to show correspondences of the type described in Equation 4 (which are stated as conjectures in [28]).

Thus far I have reviewed the *ASPIC+* framework that provides guidelines for dialectical characterisations of non-monotonic inference that are rational, and preliminary work extending *ASPIC+* so as to also accommodate arguments that express preferences and reason about values, and the subsequent generalisation of this work to define formal models of dialogues in which agents can jointly reason and engage in epistemic and practical reasoning, while also resolving disagreements about the preferences and val-

[9] Including Preferred Subtheories (see [32]), and Prioritised Default Logic (see [50])

ues that may bear on these kinds of reasoning. However, the *ASPIC+* study of sufficient conditions for satisfaction of rationality postulates (as well as other studies[8, 16, 19]) assume that agents have unbounded resources[10], and that the consistency of arguments' premises is verified prior to submitting an argument. But real-world agents are resource bounded, and the Socratic move of showing that an interlocutor contradicts himself by committing to inconsistent premises (in one or a number of arguments) is a ubiquitous feature of dialectical reasoning. We therefore need to address the challenge of developing rational resource bounded models of argumentation and dialogue that formalise real world modes of dialectical reasoning. In this regard, [14] represents a first step toward achieving this aim, in the context of classical logic argumentation [19]. The essential insight in [14], is a refinement of the ontology of arguments so as to account for the dialectical character of argumentative characterisations of non-monotonic inference. In particular, an argument's supporting formulae are distinguished according to whether they are premises assumed true, or supposed true for the sake of argument, where the latter may refer to the premises of an interlocutor.

As well as the above mentioned Socratic move, other modes and features of real-world dialectical reasoning need to be formally modelled in dialogues. For example the use of enthymemes (arguments with some missing logical structure and/or internal components) [6, 21] and relations other than binary attacks between arguments, such as *support* relations and *collective attacks* from multiple arguments. Indeed, [27] argues that these additional relation types are characteristic of human dialogue, and that they are essentially abstractions of logical relationships between constituents of arguments. The contents of these arguments when included in a commitment store, only then need yield arguments related by binary attacks. For example, that X supports Y indicates that X concludes some α that is a premise or intermediate conclusion in Y. Hence, when the interlocutors' assertional commitments instantiate an AF, it suffices that one construct an argument Y' that is the argument Y 'extended backwards' with X, on what is now an intermediate conclusion α in Y'. Moreover, [27] argues that in real world dialogues, abstract relations amongst arguments typically encode declarative information that is not explicitly communicated due to the use of enthymemes. For example, suppose Paul argues that "Tony Blair is no longer a public figure (α), and the information about his affair is not in the public interest (β), so the information should not be published (γ)" (X). Trevor counter-argues with "but Blair is UN envoy for the Middle East" (Y). The latter is an example of an 'indirect' speech act [41], where it is inferred from the use of "but", that Y is an enthymeme that attacks X, and the missing structure in Y is a rule of the form "if Blair is UN envoy for the Middle East then δ", and δ negates either one of the premises α or β, or the conclusion γ of X. The fact that there is some uncertainty as to what δ denotes when instantiating the AF defined by the interlocutors commitments, should then instigate dialectical feedback that promotes the rationality of the dialogue, by prompting Paul to question Trevor as to what δ denotes.

Other areas for future work include studying instantiations of argument frameworks by deontic and temporal logics, so as to formalise non-monotonic normative and temporal reasoning, and argumentative formalisations of hypothetical reasoning. Another

[10] Frameworks are assumed to be instantiated by *all* arguments defined by a set of formulae Δ, which may be infinite as in classical logic argumentation [19].

topic that has recently received attention is the modelling of interlocutors' mental states and the effects of these so called 'opponent models' on strategic choices made during dialogues [20, 38]. Finally, the applications envisaged in this paper will require more concerted interdisciplinary research efforts, in particular with areas such as argument mining [23] and computational linguistics more generally, informal logic and philosophy (e.g., integration of logic based argumentation with schemes and critical questions, speech act theory, and the pragma-dialectic school of argumentation [46]), and integration with machine learning to address the symbol grounding problem [18].

6 Conclusions

The last two decades have witnessed an explosion of interest in formal argumentation theory, with researches often locating the significance of the theory in its: 1) provision of an integrative bridge between computational models of logic based reasoning and human reasoning, and 2) provision of formal underpinnings for dialogue. This paper has set out to further substantiate these claims. Firstly, while the work of Sperber and Mercier has been cited to support the need for development of artificial cognitive systems whose reasoning processes are compatible with those of humans [22], this paper has cited the argumentative theory of reasoning to support use of computational models of dialogue for enhancing the quality and scope of human reasoning. Moreover, I have motivated the use of such models in supporting the reasoning of AI agents, so that through joint reasoning with human interlocutors, AI decisions that have ethical implications are aligned with human values. This paper also outlines how argumentative characterisations of non-monotonic reasoning are generalised to dialogical models that facilitate distributed reasoning, and current work[11] that aims at further developing these models for use in the kinds of applications envisaged in this paper.

References

1. L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):197–215, 2002.
2. L. Amgoud and S. Vesic. Handling inconsistency with preference-based argumentation. In *4th International Conference on Scalable Uncertainty Management*, pages 56–69, 2010.
3. K. Atkinson, T.J.M. Bench-capon, and P. Mcburney. Computational representation of practical argument. *Synthese*, 152:2006, 2005.
4. K. Atkinson, T.J.M. Bench-Capon, and P. McBurney. Parmenides: Facilitating deliberation in democracies. *Artificial Intelligence and Law*, 14:261–275, 2006.
5. T. J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.
6. E. Black and A. Hunter. A relevance-theoretic framework for constructing and deconstructing enthymemes. *Journal of Logic and Computation*, 22(1):55–78, 2012.
7. M. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[11] Note that the work reviewed is not exhaustive in its coverage of relevant research, but, given that this position paper sets out a personal vision of the potential role of argumentation based dialogue that has informed my research, I have focussed on research involving myself in collaboration with others.

8. M. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310, 2007.
9. M. Caminada, W. Carnielli, and P. Dunne. Semi-stable semantics. *Logic and Computation*, 22(5):1207–1254, 2012.
10. C. Cayrol, S. Doutre, and J. Mengin. On decision problems related to the preferred semantics for argumentation frameworks. *Journal of Logic and Computation*, 13(3):377–403, 2003.
11. A. Claridge. What is the educational value of ward rounds? a learner and teacher perspective. *Journal of Clinical Medicine*, 11(6):558–62, 2011.
12. R. Clarke. Asimov's laws of robotics: Implications for information technology - part 1. *Computer*, 26(12):53–61, 1993.
13. R. Clarke. Asimov's laws of robotics: Implications for information technology - part 2. *Computer*, 27(1):57 – 66, 1994.
14. M. D'Agostino and S. Modgil. A rational account of classical logic argumentation for real-world agents. In *European Conference on Artificial Intelligence*, pages 141 – 149, 2016.
15. P. M. 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.
16. P.M. Dung and P.M. Thang. Closure and consistency in logic-associated argumentation. *Journal of Artificial Intelligence Research*, 49(1):79–109, 2014.
17. X. Fan and F. Toni. A general framework for sound assumption-based argumentation dialogues. *Artificial Intelligence*, 216:20 – 54, 2014.
18. M. Garnelo, K. Arulkumaran, and M. Shanahan. Towards deep symbolic reinforcement learning. *CoRR*, abs/1609.05518, 2016.
19. N. Gorogiannis and A. Hunter. Instantiating abstract argumentation with classical-logic arguments: postulates and properties. *Artificial Intelligence*, 175:1479–1497, 2011.
20. C. Hadjinikolis, Y. Siantos, S. Modgil, E. Black, and P. McBurney. Opponent modelling in persuasion dialogues. In *Proc. 23rd International Joint Conference on Artificial Intelligence*, pages 164–170, 2013.
21. S.A. Hosseini, S. Modgil, and O. Rodrigues. Enthymeme construction in dialogues using shared knowledge. In *Proc. 5th Computational Models of Argument: COMMA 2014*, pages 325–332, 2014.
22. A. Kakas and M. Loizos. Cognitive systems: Argument and cognition. *IEEE Intelligent Informatics Bulletin*, 17(1):14–20, 2016.
23. M. Lippi and P. Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25, 2016.
24. H. Mercier and H. Landemore. Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology*, 23(2):243?–258, 2012.
25. H. Mercier and D. sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–747, 2011.
26. S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9-10):901–934, 2009.
27. S. Modgil. Revisiting abstract argumentation frameworks. In *Second International Workshop on Theory and Applications of Formal Argumentation (TAFA 2013)*, pages 1–15, 2013.
28. S. Modgil. Towards a general framework for dialogues that accommodate reasoning about preferences. In *Fourth International Workshop on Theory and Applications of Formal Argumentation, TAFA 2017 (to appear)*, 2017.
29. S. Modgil and T.J.M Bench-Capon. Metalevel argumentation. *Journal of Logic and Computation*, 21(6):959–1003, 2011.
30. S. Modgil and M. Caminada. Chapter 6 : Proof theories and algorithms for abstract argumentation frameworks. In I. Rahwan and G. Simari, editors, *Argumentation in AI*, pages 105–129. Springer, 2009.

31. S. Modgil and H. Prakken. Reasoning about preferences in structured extended argumentation frameworks. In *Proc. Computational Models of Argument: COMMA 2010*, pages 347–358, 2010.
32. S. Modgil and H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.
33. S. Modgil and H. Prakken. The aspic+ framework for structured argumentation: A tutorial. *Argument and Computation*, 5(1):31–62, 2014.
34. J. H Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.
35. A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proc. 17th International Conference on Machine Learning*, ICML '00, pages 663–670, 2000.
36. H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15:1009–1040, 2005.
37. H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.
38. T. Rienstra, M. Thimm, and N. Oren. Opponent models with uncertainty for strategic argumentation. In *Proc. 23rd International Joint Conference on Artificial Intelligence*, (IJCAI '13), pages 332–338, 2013.
39. S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *CoRR*, abs/1602.03506, 2016.
40. S. Russell, T. Dietterich, E. Horvitz, B. Selman, F. Rossi, D. Hassabis, S. Legg, M. Suleyman, D. George, and S. Phoenix. Research priorities for robust and beneficial artificial intelligence: An open letter. *AI Magazine*, 36(4):3–4, 2016.
41. J. R. Searle. Speech acts. In *Cambridge University Press, Cambridge UK.*, 1962.
42. N. Soares. The value learning problem. In *Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence (IJCAI-2016)*, 2016.
43. N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky. Corrigibility. In *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop*, 2015.
44. J. Taylor, E.Yudkowsky, P. LaVictoire, and A. Critch. Alignment for advanced machine learning systems. `https://intelligence.org/2016/07/27/alignment-machine-learning/`, 2016.
45. P. Tolchinsky, S. Modgil, K. Atkinson, P. McBurney, and U. Cortes. Deliberation dialogues for reasoning about safety critical actions. *Journal of Autonomous Agents and Multi-Agent Systems*, 25:209–259, 2012.
46. F. H. van Eemeren, B. Garssen, E. C.W. Krabbe, A. Henkemans, S. Francisca, , B. Verhei, and J. H. M. Wagemans. *The Pragma-Dialectical Theory of Argumentation*, pages 517–613. Springer, Dordrecht, 2014.
47. G. A. W. Vreeswijk and H. Prakken. Credulous and sceptical argument games for preferred semantics. In *Proc. 7th European Workshop on Logic for Artificial Intelligence*, pages 239–253, 2000.
48. W. Wallach, C. Allen, and I. Smit. Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI & Society - Special Issue: Ethics and artificial agents*, 22(4):565–582, 2008.
49. D. N. Walton. *Argument Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, 1996.
50. A.P. Young, S. Modgil, and O. Rodrigues. Prioritised default logic as rational argumentation. In *Proc. 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'16)*, pages 626–634, 2016.