Argument Schemes and Dialogue Protocols: Doug Walton's legacy in artificial intelligence

Peter McBurney*
Department of Informatics, King's College London
peter.mcburney@kcl.ac.uk

Simon Parsons†
School of Computer Science, University of Lincoln
sparsons@lincoln.ac.uk

Abstract

This paper is intended to honour the memory of Douglas Walton (1942–2020), a Canadian philosopher of argumentation who died in January 2020. Walton's contributions to argumentation theory have had a very strong influence on Artificial Intelligence (AI), particularly in the design of autonomous software agents able to reason and argue with one another, and in the design of protocols to govern such interactions. In this paper, we explore two of these contributions — argumentation schemes and dialogue protocols — by discussing how they may be applied to a pressing current research challenge in AI: the automated assessment of explanations for automated decision-making systems.

Keywords: Dialogue protocol, Argument Scheme, XAI.

The authors would like to thank John Woods for the invitation to contribute to this special issue, and for his forbearance when it became clear that they had been overly optimistic in their estimation of the speed with which they could complete the task of writing this paper.

*PM wishes to thank Dylan Cope and Liz Sonenberg for many interesting conversations on the topic of explanations.

†SP acknowledges funding from EPSRC grant EP/R033722/1, and thanks Nadin Kökciyan, Quratual-ain Mahesar, Isabel Sassoon and Elizabeth Sklar for many interesting conversations on the topic of explanations.

1 Introduction

Both of us, along with many researchers in artificial intelligence (AI), especially those working on computational models of argumentation and multiagent systems, have been greatly influenced by the work of Doug Walton. At the end of this paper (see Section 5) we will say a little about this from a personal perspective, but we want to spend the bulk of this paper exploring why we think Doug's work has been so influential. In short, it is because two aspects of his work — the work on dialogue protocols, exemplified by [62], and that on argument schemes, exemplified by [61] — provide a basis[1] for a solution to some of the major problems in artificial intelligence.[2] We will illustrate this by taking one such problem — the need to provide explanations for the reasoning performed by AI systems — and showing how Doug's work provides an underpinning for a possible solution. We start with this problem, and why it has become a prominent problem.

1.1 Why explanations are necessary

The third edition[3] of Russell and Norvig's "Artificial Intelligence: A Modern Approach", published in 2009, includes a history of AI from its birth (which they date to 1956, at the Dartmouth workshop, though acknowledging that work on AI was done before this point) to the time of writing. The period from 2001 is headed "The availability of very large datasets", and points to the ability of systems bootstrap from large collections of data as possibly leading to AI systems that no longer need the careful knowledge engineering that was previously necessary. The subsequent decade has seen this prediction, if not borne out[4], at least extensively tested, with impressive results on a range of applications.

Much of this success has been due to techniques from *deep learning*, that is techniques that make use of neural networks with many layers. These methods were coming into their own while Russell and Norvig were putting the third edition

[1] The use of the indefinite article is deliberate here. There are undoubtedly other solutions which would have other bases. However, that does not undermine the importance of that based on Doug's work.

[2] At the end of writing this paper we discovered another tribute to Doug Walton that focuses on the same two of his contributions, this time in the area of AI and law, namely [7].

[3] Though a fourth edition was published in early 2020, it is not yet easily available in the UK at the time of writing.

[4] It is noteworthy that much of the recent cutting-edge work on machine learning has been looking at ways to incorporate engineered knowledge into the learning process, suggesting that researchers in machine learning are beginning to feel that there are limitations to the idea of extracting all that is needed to solve every problem directly from data.

together[5], and have come to dominate work on machine learning and AI. Indeed, for many outside the field of AI, and a good number of those within who have graduated in the last few years, machine learning *is* AI, and the only kind of machine learning worth considering is deep learning. While the performance of deep learning systems is extremely impressive, there are a number of (well-known) issues that widespread use of such systems raises. Chief among these[6] is the fact that it is frequently obscure *why* a deep model gives a specific answer. This is in contrast to earlier AI methods — for example the rule-based methods of expert systems, or the causal probabilistic networks that led to the previous wave of AI applications — where it is straightforward to extract a trace of the reasoning that led to a conclusion and one could pose "what if?" questions about related situations. It is in contrast to other machine learning methods, for example decision trees, where structural information about a domain can be extracted from the model that has been learnt.

The reason that this is significant is because, as AI applications become more widespread, there will be an increasing need to be able to explain not just *what* decisions were reached, but *how* those decisions were reached. In other words, there is a requirement for AI to be *explainable*. This requirement is driven by regulatory pressure. For example, GDPR[7] regulation in the EU, requires that organisations that use AI systems to make decisions

> shall implement suitable measures to safeguard [the subject of those decision]'s rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the [organisation making the decision], to express his or her point of view and to contest the decision.

This is widely understood to mean that decisions made by those AI systems must be such that that can be explained to the subject of those decisions, since how

[5]In [26], three of the pioneers of deep learning date the breakthrough in such methods to 2009 for speech recognition and 2012 for image processing.

[6]Two others, in passing, are the following. (1) the fact that deep learning not only benefits from huge amounts of data, but *requires* it. As a result, if you work in a domain that does not have tens of thousands of examples that your system can learn from, you will not be able to create robust models. Unfortunately, areas like medicine fall into this category. Another example is the creation of software for control of autonomous ships, where there is a severe lack of publicly-accessible data on collisions. Nowadays there are very few collisions between large ships; there are many more near-collisions, but most of these are not reported outside the companies involved. (2) training deep models uses a large amount of power, and since the methodology for learning the hyper-parameters that determine whether or not a particular model is effective is basically brute-force search, training a good model is very energy inefficient. In a climate emergency, one might question the morality of widespread use of deep learning.

[7]`https://gdpr-info.eu/`

else would that subject be able to express their views and contest the result in any meaningful way?

Similarly, the European Union's Markets in Financial Instruments Directive II (MiFID II[8]), which came into force in January 2018, requires companies which provide financial information or services in which wholly automated decision have material impacts on individuals or on small and medium-sized enterprises to provide those impacted with human-understandable explanations of how the automated decisions have been made[9]. Indeed, the policy statement from the European Commission to the European Parliament relating to AI (published in April 2018) emphasizes Explainable AI as a key area of research and innovation for the next EU Multiannual Financial Framework (2021–2027), along with the areas of unsupervised machine learning and energy and data efficiency.[10]

These regulatory pressures are also present elsewhere in the world. For example, in January 2019 the Personal Data Protection Commission (PDPC), a Government agency in Singapore, released a draft model framework for the Governance of AI systems in large organizations and enterprises [1]. After public consultation during 2019, a revised version was released in January 2020. The framework is a voluntary collection of ethical principles and governance considerations that are recommended by the PDPC for adoption by organizations; the Framework is not legally binding. The Model Framework proposes two high-level guiding principles for design and deployment of AI applications:

- Organizations using AI in decision-making should ensure that the decision-making process is explainable, transparent and fair; and
- Applications of AI should be human-centric.

The Singapore Model Framework also provides guidance on when and how applications of AI should incorporate human involvement in decision-making processes.

This section has discussed the pressures from Governments and industry regulators on adopters of automated decision-making systems to ensure that these systems explain their decisions. Another pressure will likely come from the legal system. If a

[8] `https://www.esma.europa.eu/policy-rules/mifid-ii-and-mifir/`

[9] Legal or regulatory requirements to provide explanations for decisions reached by automated systems have led some people to propose inserting a dumb human into the decision process so that the process no longer appears completely automated. However, if the human only ever approved the decisions and never rejected them, then it is unlikely that European courts would accept such gaming of these regulations.

[10] `https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe`

human car driver is faced with an untenable choice, for example driving straight on and thereby hitting an oncoming car or swerving off the road and hitting a pedestrian, and if there is a subsequent legal case, judges and juries may well accept (as they do now) an explanation from the driver along the lines of, *"I was faced with an impossible choice, and in the heat of the moment I chose one way rather than the other."*.

However, if the car in question is an autonomous vehicle, that response will most likely not be acceptable to courts. Instead, courts will want to ask how that trade-off was made by the vehicle in that moment. Was it pre-coded? If so, how did the software developers make that pre-coded decision? If not, how did the software developers allow the machine to decide itself between the two options (e.g., did it make a random choice?). Courts may well also probe what ethical considerations the developers considered before coding the vehicle. What ethical training had they had before considering any ethical issues? What directives or ethical advice, and from whom, had they received beforehand? Etc. Such probative questioning by courts will not stop at the first response as with a human driver. Hence, we expect the legal system's response to cases concerning accidents involving autonomous vehicles to add further pressures on developers of AI systems to provide explanations of the decisions made or recommended by those systems.

1.2 Fairness and explanation

Note that this desire for AI systems to be explainable, is related to concerns about the *fairness* of AI and, more broadly, what is known as *algorithmic decision making*[11].

[11]The term "algorithmic decision making" is used to refer to situations in which decisions are made by a system that involves software with no human oversight or involvement. Clearly decisions made by a software system that uses AI and which has no human oversight or involvement are a subset of those reached using algorithmic decision making. In our view, "algorithmic decision making" is a bad piece of terminology, since it is perfectly possible for a human to follow an algorithm as part of making a decision in such a way that they exercise no free will, making the decision determined purely by the process encoded in the algorithm. In other words, the use of the term "algoithmic" does not imply the use of software, or the exclusion of a human from the process. One of us (SP) remembers making decisions in exactly this way when he worked in a temporary position at a Job Centre in the summer of 1988. One of the parts of the job was reviewing the record of people receiving unemployment benefit and, provided that they met some criteria to do with the length of time they had been out of work, inviting them for an interview. ("Inviting", in this case, meaning "threatening them with a loss of benefits if they did not attend".) The process was as mechanical as described — we were not allowed to exercise judgement, and what we did could easily have been carried out by software. We suspect that the reason that such a poor term as "algorithmic decision making" has come into use is a combination of its euphony (much better than "software decision making" or "computerised decision making") and the fact that many people do not know the difference between an algorithm (the process itself) and its implementation.

The concern is that whenever software is used with no human intervention, there is the possibility for it to produce results that are biased, in the sense of discriminating against individuals. Of course decisions involving humans can also be biased if the humans are biased, but part of the concern with software decisions is that they can be unscrutable (and so be hard to identify and rectify) and that they can exist even when the software designers and deployers have no intention of being biased[12].

Two well-known cases are the admissions process for students at St George's Hospital Medical School in London, and the COMPAS recidivism risk calculator. In the case of St George's [28], the medical school created a piece of software to screen applications for places to train to be a doctor. There were two aims. First, they wanted to ensure that all applicants were treated the same, something that can clearly not be the case when decisions are reached by humans (especially when the decisions are distributed across a group). Second, they wanted to reduce the load on their staff. The medical school was heavily over-subscribed (with 12 applicants for each place in 1988), and the idea was to have the software screen out some applicants so that the admissions team had less applications to consider. The system was carefully designed and then tuned until it had close agreement with the manual process. Unfortunately, the manual process was itself flawed, and the software system was found to be discriminatory, with an investigation by the Commission for Racial Equality finding that:

> as many as 60 applicants each year among 2000 may have been refused an interview purely because of their sex or racial origin. [28]

COMPAS, is a software system developed by Northpointe Inc. to help assess the risk that, on the basis of their history, an individual would reoffend. The performance of the system was analysed by ProPublica [4, 25] and found to exhibit racial bias. The analysis considered more than 10,000 real cases from one county in Florida, and compared the rate of recidivism predicted by the COMPAS software against what the individual actually did in the next two years. The headline finding was that:

> Black defendants were often predicted to be at a higher risk of recidivism than they actually were. Our analysis found that black defendants who did not recidivate over a two-year period were nearly twice as likely to

[12]Note that it is possible for bias to exist not only in the data used for training purposes or as inputs to some software analysis procedure, but even in the underlying conceptual abstractions that allow the data to be recognized as *data* and thus enable its collection; for an example, see [19]. Econometricians analyzing national accounts data face similar issues, for example, when the definition of employment ignores unpaid work done by family members within households or on farms.

> be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent). [25]

In both these cases, the software system making decisions does so in a way that is biased. In the case of the admissions system, the software was designed to replicate an existing decision process that was already biased. In the case of the recidivism system, the designers apparently [3] tuned the system to ensure that its accuracy was the same for both black and white individuals — they assumed that doing this would make its decision fair. However, as above this turned out not to be the case for some reasonable definitions of "fair", in particular the one alluded to in the quotation above, that the rate at which defendants were wrongly classified as higher risk should be the same regardless of whether the defendant was black or white. Subsequent analysis [3] has shown that it is impossible for both these notions of fairness — that the accuracy of predictions do not vary by race, and that there is no disparity in incorrect misclassification as higher risk — to be simultaneously satisfied. Indeed, aiming for equal accuracy of predictions leads directly to a disparity in misclassification to a higher risk category. Such concerns about the fairness of AI lead back to the desire for AI to be explainable because if one can check the reasoning that an AI system uses, then it will be possible to check that reasoning for bias [64].[13]

1.3 Explainable AI

The last few years has seen a surge in work on explainable AI, or XAI. Much of this work has centred around creating explanations for machine learning models, especially those that look to many users like "black boxes", in other words inscrutable oracles that are inherently impossible for people to understand. A typical approach is to take a black box model and train another model that is easier to understand on the same data, and use that second model to explain the decisions made by the first. This is the thrust of [16], which creates an ensemble of decision trees as an explanation of a, more complex, deep neural network model. Another take on the same issue is to explain a decision by plotting out the local area around the point where the decision needs to be made and creating a model of that [43]. The intuition here is that the inscrutability of models — that they consist of complex multi-dimensional surfaces separating different outcomes — will often not exist at a local level, allowing simple, and hence easy to understand, rules to be identified that explain the decision. A criticism of this work, and much of the other efforts in XAI is

[13] Of course, this is not the only way to ensure fairness, and much of the work on the fairness of AI systems does not attempt to do this through explainability.

that they are developed by the same people who build the black box models in the first place, start from the thing to be explained, and create a solution by simplifying it. This is a process that takes very little account of what the people who want the explanations would find helpful [39].

Miller's [39] examination of the literature on explanation, follows [23], among others, in suggesting that many explanations presented by people focus on describing the underlying causal mechanisms, and, further [24], that these explanations are presented in the form of a conversation. As [39] discusses, [5] goes further in suggesting that explanations are presented not just as conversation, but as *arguments*, in the sense of the provision of justifications for the assertions that are made. The research in [5] is drawn from the analysis of a number of explanations from human conversations — that is where one person explains something to another. Given this, admittedly rather limited[14], evidence, it seems plausible that an argumentation-based approach to explanation will be a promising approach for adoption by AI systems. Below, we sketch some requirements for computer-based explanations, giving a first-principles analysis to complement the discussion above, and the point to ways in which Doug Walton's work can be used to underpin these requirements.

2 Asking for and assessing explanations

2.1 What do we need for computer-based explanations?

What would we require in order to have automated explanations? A first requirement — and challenge — would be to generated explanations automatically for AI decision systems. As mentioned above, for some types of AI systems, such rule-based expert systems and causal probabilistic networks, automatically generating explanations is straightforward, by generating a trace of the reasoning undertaken by the system in reaching a conclusion.

For other types of AI systems, especially those which operate at a low level of granularity, such as image classification programs analyzing individual pixels and their neighbourhoods, this is not necessarily at all straightforward. Why automated generation of explanations for such systems is difficult is because the level of operation of the AI system is at lower level of the objects being classified than any level containing human meaning. For a human being recognizing images of faces for example, parts of the face are arguably very important to recognition and classification, for example, colour of hair, shape of hairline, size of ears, presence or absence

[14] As [5] explain, their analysis is based on 30 examples of explanation, but they are from a single conversation, itself taken from [51]

of a beard, etc.[15] Such parts have human meaning and can be readily described to other humans as the reason for a particular classification. If, instead, an AI system uses lower-level elements of images, such as individual pixels, or the relationships between nearby pixels (eg, identifying edges by means of observed differences in pixel colours) for facial recognition or classification, then these lower levels will typically have no human-understandable meaning. It is generally not obvious how the use of such lower-level elements could be aggregated or assembled automatically into a higher-level explanation able to be understood by a human. Thus, automated generation of explanations is difficult challenge for these types of AI systems.

In this paper, however, we will ignore the challenges involved in the generation of explanations. Our focus will be on assessment of an explanation that has somehow been produced, by automated means or manually. Given that an explanation has been created, what is needed for its automated assessment by some entity seeking to obtain an explanation for a decision of an AI system? Based on human-to-human explanations, we might expect any machine assessment to have several features.

The first feature is a means for the formal representation of explanations, where by *"formal"*, we mean machine-readable. This is necessary for automated parsing of the explanation, as the first stage in a process of automated analysis and assessment, and possibly also automated comparison with alternative arguments. As mentioned above, this process is well-known and straightforward in cases where explanations may be constructed from sequences of syllogistic or mathematical deduction (as in rule-based Expert Systems) or from sequences of causal influences between time-ordered events (as in Bayesian Belief Networks). Automated parsing and reasoning over such explanations is routine in AI and in Computer Science[16]. However, there are many other types of inference besides logical deduction and other types of explanation besides sequences of causes and effects. It behooves us therefore to seek more general formalisms for representing explanations.

2.2 The role of argument schemes

One such generalization are argumentation schemes with critical questions. Doug Walton was a pioneer in the study of argumentation schemes, both individual

[15]This account of how humans recognize faces differs from that given in Oliver Sack's book, *"The Man Who Mistook His Wife for a Hat"* [46]. When one of us (PM) wrote to Sacks in 1987 to contest his account and to propose an alternative, Sacks replied with a suggestion for an experiment to decide between the two alternative explanations. Only decades later did PM learn that Sacks suffered from prosopagnosia.

[16]For instance, every version of Microsoft's Windows Operating System since the release in 1995 of *Windows95*, has included a Bayesian Belief Network for the diagnosis of the likely causes of printer faults.

schemes and collectively.[17] He told one of us (PM) that he had been led to consider these schemes for pedagogical reasons — to make it easier for his students to recognize and critically analyze informal arguments. Only later did he realize that their study could have theoretical and practical implications. His 1996 book [54] appears to have been his first work looking at multiple schemes, but he had written earlier books on particular types of informal argument, for example, on *Ad Hominem* arguments [52] and Slippery Slope Arguments [53].

Arguments schemes are a form of default reasoning where a claim is posited as presumptively true or to be endorsed by default. A rational reaction to the claim may investigate the assumptions being made, implicitly or explicitly, in endorsing the claim and assess whether or not these assumptions hold in any particular case. We could consider consideration of the assumptions to be an assessment of the validity of application of the scheme in a particular case. Endorsing a claim (especially a claim proposing that an action be executed) may entail commitments to endorsements of other claims or to other actions. Arguably, a rational decision–maker (one making decisions based on reasoned grounds) would therefore only endorse the default claim both knowing these commitments and *taking any decision under advisement*, i.e., informed by that knowledge. Thus, a rational decision–maker would also assess the commitments that endorsement of a default claim would entail. As well as eliciting the assumptions behind a presumptive conclusion, critical questions can explore the existence and nature of such entailments.

Argumentation schemes with their associated critical questions have found application in AI, for example in the development of automated argument in practical reasoning [8], in automated dialogues over commands [9], and in automated selection of statistical models for data analysis [48]. Many argument schemes involve default conclusions which are logically fallacious, and so their study has been undertaken in that branch of argumentation theory known as Informal Logic. Even though logically fallacious they may plan an important role in society, particularly in situations where information is incomplete, inconsistent or uncertain. As an example, *Ad Hominem* arguments are criticized by most scientists, since they appear draw conclusions about the content of an argument from personal attributes of the proponent of that argument. Science, it is often argued, should be an objective activity, and so *Ad Hominem* arguments are typically disparaged by scientists. Yet, these arguments play a great role in legal proceedings, because they allow the court to assess the testimony of witnesses and of experts. Over time, in most legal jurisdictions, rules have developed as to when and how such arguments may be made in considering testimony.

[17] See [30] for a history of argumentation schemes and related forms of reasoning.

Another example are *epideictic* arguments, which involve drawing conclusions about the substance of claim from the form of its presentation. Although clearly logically fallacious, there are circumstances where this form of reasoning is rational, as William Rehg has argued [42]. Indeed, there are circumstances where epideictic reasoning is also commonplace, as in assessments made by venture capitalists of potential investment proposals from start-ups. In this situation, potential investors may have little past experience on which to base an investment decision, and the start-up may face an uncertain and fast-changing business environment. The marketing plans and financial forecasts of the start-up management team will almost certainly not prove accurate, and so the team's ability to modify their plans in the light of operational experience becomes a better indicator of their potential success than the contents of the current plans themselves. Such abilities may best be assessed, not by the written plans and forecasts, but by the management team's ability to respond to probative questioning from the venture capitalist.

Not only is the use of such logically fallacious informal arguments widespread, there is a strong argument that modern society could not function without their use. Philosopher Charles Willard, for instance, has argued [65] that in a society which depends on complex technology that is too vast and changing too quickly for any one person, or even a small group of people, to ever master completely, then we all need to rely on arguments from authority and on assertions made by experts. The COVID-19 pandemic[18] that so occupies our current attentions illustrates our society's reliance on such arguments with great immediacy. The point is not to avoid such a reliance, because that is infeasible, but rather to make our reliance as rationally justified as possible (within the time available in each case) by means of rational interrogation of the claims of authorities and of experts, and of their supporting arguments and sub-claims.

Thus, for instance, in the case of COVID-19, many governments have relied on advice from expert epidemiologists. Given a particular claim from a particular expert epidemiologist, we could interrogate it according to the Argument Scheme from Expert Opinion that Walton articulated and studied in [55, page 210]. This scheme was presented as an argument with two premises and a default claim, along with six critical questions. Using the notation of a later presentation[19] of the argument

[18] For the benefit of any readers who were born after the pandemic, particularly if it has largely disappeared from the historical record in the interim, we note that this paper was written in the throes of the second wave, and that the pandemic as a whole greatly disrupted all aspects of life across the world, including the writing of this paper.

[19] As a trivial example of the wide-ranging nature of the impact of the pandemic is the fact that one author (PM) owns a copy of [55] but has not been to his university office, where the book is located, for 8 months.

scheme [60], E is an expert in some field of knowledge F comprising a finite collection of propositions. The argument scheme consists of:

1. **Major Premise:** Source E is an expert in field F containing proposition A.
2. **Minor Premise:** E asserts that proposition A in field F is true (false).
3. **Conclusion:** Proposition A may plausibly be taken to be true (false).

Walton [55, page 223] proposes six critical questions for this scheme, labelled **CQ1** through **CQ6**, as follows:

CQ1. Expertise Question: How knowledgeable is E as an expert source?

CQ2. Field Question: Is E an expert in the field F that A is in?

CQ3. Opinion Question: What did E assert that implies A?

CQ4. Trustworthiness Question: Is E personally reliable as a source?

CQ5. Consistency Question: Is A consistent with what other experts assert?

CQ6. Evidence Question: Is E's assertion based on evidence?

To these six questions, we would add another:

CQ7. Self-interest Question: Is it the case that E does not stand to gain by our endorsement of proposition A?

As a simple example of the use of this scheme, consider that Anthony asserts that wearing a mask is an effective way to limit the spread of the Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that causes COVID-19. Considered through the lens of the Argument Scheme from Expert Opinion, we might want to check that we can provide positive answers to the critical questions before we are prepared to accept Anthony's argument. In this case we can accept the argument, since (CQ1) Anthony is a extremely knowledgeable, having been extensively cited; (CQ2) Anthony is an expert in a relevant field, that of infectious diseases; (CQ3) Anthony made assertions in [27] implying that wearing a mask was an effective way to limit the spread of SARS-CoV-2; (CQ4) we have no knowledge of Anthony lying, so can consider him a trustworthy source; (CQ5) his advice is consistent with what other experts, for example those in the World Health Organisation[20]; (CQ6) Anthony's

[20] `https://apps.who.int/iris/handle/10665/332293`

assertion is backed by evidence, listed in [27]; and (CQ7) there is nothing to suggest that Anthony has anything to gain by our endorsement of his claim that wearing a mask is an effective way to limit the spread SARS-CoV-2.

Note that in order to accept Anthony's claim, we need to examine *all* of the critical questions. If we cannot give a positive answer to any one of the questions, the conclusion should not be accepted. For example, consider Donald, who makes the opposing claim to Anthony, that wearing masks is not helpful in the context of the COVID-19 pandemic. Even if one accepts that Donald is an expert in the field, the fact (CQ4) that he is known to have repeatedly lied on the matters related to the pandemic[21] means that we cannot answer the "trustworthiness question" in the affirmative, and hence Donald's argument claim cannot be accepted.

In the above example, there is only level of analysis. We took the argument, and applied critical questions to that argument. However, a multilevel analysis may sometimes be appropriate. Consider Neil, who claims, for example, that during the pandemic, no more than six people should gather together indoors to limit the spread of the disease. As an epidemiologist, Neil is an expert on disease transmission, and when asked for evidence to support his argument (CQ6), would point to the computational diffusion model that generated the results. In other words, the claim about the "rule of six" rests on the output of a computational model.

Why should we accept that output? Well, the epidemiologists who developed the model would claim, in effect, that it is an oracle which, much like an expert, considers a range of factors that are outside the grasp of most humans. The oracle weighs these factors and produces a summary that the non-experts can use to guide their behaviour. Since the computer model is treated as an expert, we might consider the evidence that it produces for Neil's claim in the same way that we consider the claim itself, that is as an instantiation of an Argument from Expert Opinion. If we do this, then we might want to subject it to a second level of analysis, to check whether neil is justified in relying on it. If we do so, then, in order to answer the "trustworthiness question" (CQ4) it might be wise to ask the opinion of professional programmers, another group of experts part of whose expertise is the ability to establish whether software is reliable, that is whether the outputs of that software are trustworthy.

When experienced programmers look at models like Neil's, they usually find they were built incrementally and with very poor or no software engineering practices. That is, there is no or little documentation, no standard good development models, no agreed statement of specifications, no formal design, no rigorous testing of the

[21] See [11] for a list of Donald's many lies on the subject between the start of the pandemic and November 2nd 2020, and [63] for a record of his lies as President. As of September 3rd 2020, the number of lies that Donald had told since taking office was more than 22,000 over the course of 1320 days.

components, and no independent testing by professionals other than the programmers who built the model.[22] This might lead us to question the relaibility of the model, and hence whether Neil's original claim holds.

In contrast to the situation in computational epidemiology, some disciplines which regularly use simulation models, such as economics, specialist expertise increasingly exists on how to evaluate such models, for example [32, 38]. Thus we see an instance of Willard's argument on the inter-connected complexity of contemporary life: evaluation of a statement about the best public policy to reduce the risk of infection during the pandemic may require, for its resolution, evaluation of claims about particular computational simulation models in epidemiology, which, in turn, may require evaluation of claims about software engineering best practice and their application to the particular epidemiological model; few people if any have the necessary skills to evaluate all these claims across the different disciplines involved, from public policy to epidemiology to simulation modeling to software engineering.

2.3 The role of dialogue

A second feature is that evaluation and assessment of explanations might best be undertaken within the context of a dialogue, between an *explainer*, either the entity which generated the explanation or an entity able and willing to answer questions about the explanation, and an *explainee*, an entity seeking to assess the explanation, For human interactions, if one person seeks from another person an explanation of something, and the two have an appropriate social relationship allowing them to engage in a conversation of equals,[23] then our contemporary western cultural experience would lead most of us to expect the two entities to engage in a dialogue involving questions and responses about the explanation. We do not call these responses *"answers"* because they may not be intended by the responder to be answers to a prior question and because, even when so intended, they may not satisfy the questioner.

The questions may serve a number of purposes: they may seek further clarifica-

[22] As an example of such analysis, see the anonymous critique a software developer of the code of the Imperial College COVID-19 epidemiological model by published on the Web in May 2020 at: `https://lockdownsceptics.org/code-review-of-fergusons-model/`

[23] Habermas [20] discusses such social relationships. In this sub-section, we are ignoring interactions which are normally adversarial, such as criminal and military interrogations or courtroom cross-examinations. One of the dialogue types which Walton and Krabbe include in [62] is Eristic dialogues, which are adversarial interactions where one or both parties give vent to anger or frustration. Even these dialogues have been studied by argumentation theorists, e.g., by Dov Gabbay and John Woods [17, 18]; this work has potential applications, for instance, in customer service centre operations.

tion of the explanation; they may seek clarification regarding a response to a prior question; they may seek to identify or make explicit any underlying assumptions in the explanation or in the responses; they may seek to identify consequences of the explanation or of the response (for example commitments to particular beliefs or actions entailed by endorsing the explanation or a response); they may seek to expose internal inconsistencies in the explanation, or in the responses, or in both explanation and responses when considered together; they may seek to contest or argue with the explanation, or its assumptions or consequences, or those of the responses; and, as with any linguistic interaction between two or more parties, the questions posed may seek to clarify previous utterances or concern the operation of the dialogue itself, for example, if there is sufficient time to ask further questions. In other words, this conversation between some person or machine seeking an explanation generated for an AI system and a person or machine who has proposed such an explanation could easily take the form of a dialogues involving questions and responses. For convenience in this paper, let us call these Explanation-Question-Response (EQR) dialogues.

To enable machines to automatically engage in such EQR dialogues, we need to define the rules of the dialogue — their formal syntax, their semantics, and their pragmatics. Although these terms are taken from linguistics, over time they have come to have subtly different connotations in disciplines other than linguistics, firstly in mathematical logic, and then in computer science and AI. In particular, as we discuss in [35], for autonomous computational agents engaged in dialogic interactions, a formal semantics is needed for the agents (and their human or machine designers) to be able to verify, as best they can, that different agents engaged in a dialogue share the same understanding of each other's utterances and of the dialogue itself.[24] Moreover, having a formal semantics and pragmatics for utterances and dialogues can greatly facilitate (or hinder) the computational implementation of interactions. In [35] we discuss these issues at length; here we will briefly mention each element with respect to EQR dialogues.

Syntax

The rules of syntax for a computational dialogue typically govern the permitted forms of utterances and the rules applying to their use. An agent communications language such as ACL developed by the Foundation for Intelligent Physical Agents FIPA (now IEEE FIPA) [15], for example, specifies very strictly the form of each of the 22 permitted utterances, although it has no rules or protocols regarding their

[24] Michael Wooldridge showed in [66] that a sufficiently clever software agent can always present to an external observer an insincere representation of its own internal state.

combination. Computer scientists attempting to use this language for agent communications quickly realized that more structure was needed, and so developed specific interaction protocols, for example, for running Dutch auctions [14]. Such protocols, although well and good for their particular intended purposes, lack generality. What was needed was a general theory of dialogue which allowed for different types and purposes of dialogues.

This was found in Doug Walton's 1995 book with Erik Krabbe, *"Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning"* [62]. This work presented a classification of human dialogues in terms of three dimensions:

- What the participants each knew before the dialogue commenced;
- What each participant intended to achieve by participating in the the dialogue (i.e., the goals of each participant); and
- What the goals of the dialogue are.

With these dimensions, Walton and Krabbe identified and analyzed six types of dialogue: Information-Seeking, Inquiry, Persuasion, Negotiation, Deliberation and Eristic. This classification and these dialogue types have been quite influential within AI with computational models being proposed for each of these types (see [35] for a review of applications). Walton and Krabbe do not claim their list is comprehensive, and indeed other types have been studied by researchers in AI. In earlier work [34, 35], we presented a list of the key elements needed for specifying the syntax rules of a dialogue between computational agents, drawing both on Speech Act theory from the Philosophy of Argumentation (as does the ACL language of IEEE FIPA) and on Walton and Krabbe's classification in [62].

Although very influential in AI, the Walton and Krabbe classification is not without some challenges. In a context of autonomous agents, one would have to ask how a dialogue type, an entity without agency, could have goals. At best, *"the goals of the dialogue"* might be understood as the maximal subset of shared goals of the participants for their participation in the interaction, but that would assume they share any goals. In a multi-agent context, that assumption may not apply. In other work, one of us identified *"the goals of the dialogue"* with the set of possible outcomes of interactions conducted under the rules of that dialogue and used these sets of outcomes to design an efficient means of storage of dialogue types [40].

Moreover, in any computational system where participating agents may be designed by independent teams of software developers, there is no guarantee that the stated goals of each participating agent are in fact their real goals. Even without any insincerity on the part of the participants of their design teams, software agents

may have buggy code, and so may act contrary to their stated goals.[25] For example, a participant in a dialogue may wish never to a reach a conclusion, or may wish to delay reaching a conclusion until after some other event has occurred, or may join an interaction in order to delay or distract another participant, or just to cause confusion.[26]

For EQR dialogues, we could conceive the appropriate dialogue type to be a Persuasion dialogue, where the Explainer is trying to persuade the Explainee to accept or endorse the explanation provided by the Explainer. The incorporation of critical questions, however, may lead us to consider these interactions as Information-Seeking dialogues (where Explainee is seeking an explanation from Explainer) or Information-Giving dialogues (where Explainer is providing information in the form of an explanation about the operations of some AI system to Explainee). Information-Giving dialogues are not analyzed in the Walton and Krabbe typology [62]. However, in may applications of EQR dialogues, Explainee may wish to see how well and by what means Explainer is able to convey an understanding of the operations of the AI system in this particular case, for this particular decision, and so the dialogue may be closer in form to the Query dialogues of [12], where questioner wants to hear and understand, not just a claim itself, but the arguments for the claim.

Semantics

As far as we are aware, Charles Hamblin was the first person to present a semantics for question–response interactions, in his 1957 PhD thesis [21]. Hamblin's semantics was based on alternative possible worlds, with different responses corresponding to certain propositions being true in different possible worlds.[27] Hamblin later expanded these ideas in a paper that became well-known in linguistics [22]. The subject of the semantics of questions and of question–response interactions has since become a topic of great interest in theoretical linguistics, and there are now several alternative theories; see Floris Roelofsen's linguistics encyclopedia entry [45] for a recent review.

[25] For the same reason, the consoling assumption of mainstream economists that agents always act in their own self-interest cannot be made by computer scientists.

[26] Some of these disruptive behaviours have been observed in industry-wide discussions over new computer standards [37].

[27] Hamblin's PhD, which was submitted in 1956, included one of the earliest instances of possible worlds semantics, alongside those of Richard Montague (initially in 1955), Carew Meredith and Arthur Prior (1956), Stig Kanger (1957), A. Bayart (1958, 1959), Saul Kripke (1959, 1962) and Jaako Hintikka (1962). See [13] for a partial history of possible worlds semantics. Hamblin had been a student of Karl Popper and Hamblin's own student Jim MacKenzie argues in [31] that Hamblin was strongly influenced by the ideas of both Popper and Wittgenstein.

In this paper we are proposing the use of argumentation schemes and critical questions for modeling arguments and questions in EQR dialogues. The various semantics for question–response interactions explored in linguistics do not formally incorporate the structure of argumentation schemes and critical questions. The critical questions are not randomly asked, but are specific to the presumptive claim of a specific argumentation scheme, and to its specific (albeit possible implicit) assumptions and its specific potential consequences. We believe this argumentation theoretic structure is important for understanding (and thus for modeling and automatically generating) the reasons why particular questions are asked and for the overall structure of the EQR dialogue in which the questions sit. The semantic frameworks found in linguistics, because they are not based on an explicit argumentation theory, seem too coarse for this purpose.

As an example of how a computational semantic structure can incorporate an explicit philosophy of argument based on argumentation schemes, we mention the work of Katie Atkinson and Trevor Bench-Capon in [6]. Their approach used a framework based on the Alternating-Time Temporal Logic of [2] to create a formal semantics for the syntax for multi-party practical reasoning presented earlier in [8]. The last-cited work articulated a framework for dialogues over what actions to take in some situation (i.e., practical reasoning dialogues) building on Walton's Argumentation Scheme for Practical Reasoning in [54]. We believe that a similar approach would be fruitful for EQR dialogues.

Pragmatics

The pragmatics of utterances and dialogues concerns not their form (the syntax), nor their relationship to truth or reality (their semantics), but other aspects of their meaning unrelated to truth. The most common aspect of meaning unrelated to truth concerns how and when utterances are used, for example: what pre-conditions apply to their use, and what consequences usually follow from their use. In the English language, for instance, asking *"Do you have the time?"* normally results not in an affirmative *"Yes"* response if the responder has the time, but in the provision of the time itself. So part of the meaning of this question is the fact that responders to the question usually answer another question, *"What is the time?"*

It makes sense to talk about the pragmatics of dialogues as well as of utterances, particularly when dialogues are nested, concatenated or interleaved. For example when participants in a Negotiation dialogue start to enact an Information-Seeking dialogue, one may ask if this diversion is somehow necessarily pre-determined by the first dialogue or its contents, or whether it is an appropriate diversion at this point or elsewhere in the first dialogue, etc; see [34] for a discussion of these issues.

Although Speech Act theory from the Philosophy of Language, which is focused on the pragmatics of utterances, has been very influential in the branch of AI devoted to agent communications, the computational study of pragmatics of utterances and of dialogues is still only its infancy in AI.[28] As an example of such work, our paper [36] presents a formal game-theoretic semantics for dialogues over actions, in which the semantics provides a framework for two pragmatic features of speech acts over actions: firstly, the fact that in modern western cultures, such speech often require acceptance by the intended recipient (so-called *"uptake"*) before such utterances create any action commitments; and secondly, that once a commitment is incurred, the rights of revocation of that commitment may no longer lie with with the person who made the utterance.

At first glance, uptake and revocation may be considered unimportant for EQR dialogues because these dialogues do not appear to be concerned with actions. However, insofar that explanations for decisions made or recommended by AI systems do involve actions, whether these actions are before, alongside, or subsequent to the operation of the AI system, these two considerations will be important. For instance, if future regulations or laws governing AI systems require that any implementation of an automated AI decision-system includes both an explanation of how the decision was reached for the intended subject of the decision and also an endorsement (i.e., uptake) of that explanation by the subject (acting as an explainee) before any execution of the decision, then these two pragmatic aspects will be crucially important. In the developed world we now have several decades of experience asking medical patients for their informed consent before implementing medical procedures and treatments, so modeling and implementing these aspects may well be relatively straightforward.[29]

3 The Nosenko Case

The case of Yuri Nosenko, a Soviet citizen who defected to the USA on 4 February 1964, is instructive. Nosenko arrived claiming be employed by the USSR Komitet Gosudarstvennoy Bezopasnosti (KGB) and to have first-hand knowledge of the period in which Lee Harvey Oswald, President Kennedy's assassin, spent as a defector in the USSR, including having seen his KGB files. Based on this knowledge, Nosenko claimed that the USSR had not used Oswald to assassinate Kennedy and indeed that the KGB had played no role in his death.

Opinion was strongly divided within the US Central Intelligence Agency (CIA)

[28] Arguably, it may only be in its infancy in Linguistics also.

[29] Although it is not clear that it will be; see [49] for a critique of these practices in medicine.

as to whether Nosenko was a genuine defector or a Soviet plant, intending with his defection to deceive the CIA in some way or simply to cause confusion.[30] He apparently had detailed knowledge of some aspects of KGB operations, but lacked knowledge of others (such as KGB office and human resource procedures). Over the course of the seven years following his defection, management at CIA went through periods of apparent strong belief in Nosenko's sincerity, and periods of apparent strong disbelief. In the former periods, Nosenko was treated well, given free accommodation and even given money. In the latter periods, he was held in solitary confinement and interrogated with ferocity. Among the strongest sceptics of Nosenko's sincerity was the long-term CIA Chief of Counterintelligence, James J. Angleton.[31]

Eventually, CIA leadership in 1969 officially accepted Nosenko as genuine, and he was put on the payroll as a consultant, helping to train CIA officers, for example. As late as 2007, however, Tennent Bagley, a CIA officer who had been involved in the case from the start, published a detailed account arguing for the case that Nosenko was indeed a plant [10]. Nosenko died in 2008.

A key first question for the CIA was thus whether or not to believe Nosenko was genuine. If he was genuine, then so too presumably were his claims about the files he had seen on Oswald, and the denial of Soviet involvement in the Kennedy assassination. But this first question was not the only important question. A second key question, independent of the first, was what should CIA let the Soviets believe was their (the CIA's) answer to the first question. In other words, even if CIA believed (or did not believe) Nosenko, what should they allow the KGB to know — that they did believe him or that they did not?

These two questions arise in any case of a defector, and indeed the KGB would have faced the same two questions in reverse when Oswald had defected to the USSR in 1959; likewise, the CIA would have faced them again when Oswald returned to

[30] That intelligence agencies on both sides of the Cold War sought to create confusion in their opponents is well-attested, e.g., see [44]. As an example in the reverse direction to the Nosenko case, Lukes has argued [29] that the show trial and execution of Deputy Prime Minister and former Communist Party General-Secretary Rudolf Slánský and other leading Government officials in Czechoslovakia in 1952 was facilitated by a western intelligence operation which sent false compromising letters to leading party members as part of an operation to sow confusion in Czechoslovakia. A book by journalist Stewart Steven [50] claimed that all the show trials across the region in the late 1940s and early 1950s were the result of a sophisticated western intelligence effort, called *Operation Splinter*, to cause division between the ruling communist parties in the the USSR and those in its Eastern European satellites; however, the claims of the book may be false, and the publication of the book in 1974 may itself have been a disinformation effort intended to cause confusion.

[31] A story based on the Nosenko case features in a 2006 film by Robert de Niro about the life of Angleton, *The Good Shepherd*.

the USA in 1962.[32] The answers to these questions had a special resonance in this case because of the Kennedy assassination aspect. For the CIA to lead the KGB to believe that the CIA doubted the sincerity of Nosenko would have then led the KGB to believe that the CIA doubted Nosenko's claims of no Soviet involvement in Kennedy's assassination. Even if the CIA did doubt those claims, was it in the interests of the CIA (or the USA) for the KGB to think that the CIA may consider the Soviets responsible for the assassination? While enquiries were still ongoing — the Warren Commission into the assassination only reported in September 1964 — it would have behooved the CIA to not allow a clear indication of its conclusion to the first question to be communicated to the USSR, even if a determination had been reached.

Two further complications arise here. One is that the evidence in this case, both that from questioning Nosenko and that from other information, was not clear cut.[33] If the KGB intended to sow confusion with a false defector, then these inconsistencies may well have been deliberate. On the other hand, even if not deliberate and Nosenko sincere, the KGB may also have known about the inconsistencies. Hence, if CIA wanted to convince KGB that their determination about Nosenko's sincerity was itself sincere, then they could not reach that determination (or pretend to reach that determination) too quickly or readily. In other words, the seven-year back-and-forth CIA effort to decide what to think about Nosenko may itself have been a feint, to convince the KGB that the final conclusion was reached with difficulty, and was thus itself sincere.[34] Why that would be necessary is because of the second complication: In any military conflict, it is usually very difficult to communicate a message to your enemy and have them believe it straight away; they will naturally be suspicious of any message you send them directly. For this reason, intelligence agencies may not initially reveal or expel agents of foreign powers they learn are working inside them, because such agents can be useful for the communication of messages to the enemy which the enemy are more likely to believe than direct communications.[35]

In the Nosenko case then, we have a Nosenko-explainer answering questions from

[32]The fact that the USSR accepted Oswald as a defector but sent him to the relative isolation of factory work in Minsk, may have been an indication of a lack of trust by the KGB in his sincerity. Similarly, the fact that Oswald does not appear to have faced any impediment to his return to the USA, with the US Embassy in Moscow even lending him money for the fare, despite his earlier renunciation of his US citizenship and public defection to the USSR, would have led some in the KGB to conclude that his first defection had not been genuine, i.e., that he had been a US plant (although not necessarily working for the CIA).

[33]As Bagley shows in [10].

[34]The difficulty of computational modeling of feints in human interactions is discussed in [33].

[35]Some people believe this is one reason why the UK intelligence agencies were slow to expose the Cambridge spies in the 1940s and 1950s.

a CIA-explainee. The explainer may have been seeking to deceive the explainee, and the explainee would have tried to detect such deception. Even if deception by the explainer existed and was discovered by the explainee, the explainee may not have wished to inform explainer Nosenko of this. The CIA-explainee may also have wished to deceive the USSR (specifically the KGB) about whether or not they believed the explanation given by explainer Nosenko. Hence, the explainee's actions, including the environment of the interrogations (e.g., the use of solitary confinement), the lines of questioning adopted, and the order and content of specific questions, may have been part of a larger deception effort aimed at the KGB. Even Bagley's late book [10] may have been part of some greater deception effort.

The purpose of this example is to show the difficulty of accounting for *all* relevant factors and considerations in any computational modeling of explanation dialogues. Both the explainer and explainee may have multiple objectives or agendas in which the Explanation-Question-Response dialogue plays only a small part. These objectives may be in conflict with one another, and may change in the course of the interaction. To achieve particular objectives, either or both the parties may seek to deceive the other, and to deceive external entities who are not parties to the EQR dialogue.

4 Conclusions

Alongside the recent rise to prominence of Machine Learning and Deep Learning within AI has arisen the associated challenge of automatically generating explanations for how automated decision-systems reach the conclusions they do. This challenge is driven by strong pressure from governments and industry regulators in many sectors of the economy to make automated decision-systems and recommendation-systems transparent and fair. For most model-driven AI systems, such as rule-based expert systems, generating explanations for automated decisions is relatively straightforward. For many machine learning and deep learning systems, this task is not. In either case, creating automated explanations leads to a subsequent research challenge: How may we analyze and assess these explanations, and how may we undertake this task automatically?

In this paper, we have outlined an approach to the challenge of automated assessment of explanations drawing on two areas of the philosophy of argumentation to which Doug Walton made important contributions: the study of argument schemes and their associated critical questions, and the classification of types of dialogue he developed with Erik Krabbe. Both these areas have had strong influence in Artificial Intelligence over the last quarter century, particularly in the area known as Agent

Communications. This area seeks to enable automated communications between autonomous intelligent software agents, in other words automated machine-to-machine communications.

In presenting the approach in this paper we have not considered other work of Walton's which is relevant, in particular his study of explanation dialogues, e.g., [56, 57, 58, 59]. We have also not yet considered aspects highlighted by the Nosenko case in Section 3, such as the broader intentions of the participants and the possibility of deception by either or both of Explainer and Explainee. Despite the study of lies, deception and equivocation having a long history in philosophy and theology, computational models of these phenomena are only just emerging, e.g., [47]. Applying these various elements to this challenging problem domain remains future work.

5 Memories of DW

SP: Before I sat down to write this, I thought my first memory of Doug was from the 1996 Formal and Applied Practical Reasoning (FAPR) conference in Bonn which (and this I *am* sure about) was when I first came across the Informal Logic school of work on argumentation. I, like a number of the other attendees at the conference, came to work on logic and argumentation through the AI tradition, and were wholly, and embarrassingly, unaware of this other tradition. When I started to write this, I thought I would check that Doug was there and revisit what he presented. However, I can find no record of his presence — neither in the proceedings nor in any of the material that is now online[36]. As a result, I am no longer sure whether I know Doug from FAPR, or that I became aware of his work around this time though the work of people like Chris Reed, who was quick to connect work on argumentation in AI with that from philosophy. Ultimately, though, it doesn't matter where I first met Doug. What is important, is that he became a near ubiquitous presence in my academic life (and I mean that in a good way). Very quickly Doug's work — initially that on dialogue, subsequently that on argument schemes — became pretty central to a lot of what I work on, and Doug himself turned out to attend many of the events that I went to. He was always interesting to listen to, and though I know some folk who found some of his examples to be a little, shall we say "traditional", I always found him to be both courteous and respectful of everyone I saw him interact with. He was always generous with his time, and in that, and his astonishing productivity, I have long thought of him as a role model, and will continue to do so.

[36]Of course, this was still in the dark ages pre Web-2.0, and, as a result, very little of the conference was ever published online.

PM: I first met Doug in 2000 at Pitlochry in Scotland, at the week-long *Argumentation and Computation Symposium* which Chris Reed and Tim Norman organized for philosophers of argumentation to meet computer scientists, held at Bonskeid House.[37] Doug was friendly and courteous, and – I say this as an Australian and intending it as a compliment – very Canadian. I subsequently met him frequently at various conferences and workshops and he was always the same. He was also very helpful to me in providing memories of Charles Hamblin, an Australian philosopher of argumentation and pioneer of computer science, whom he had met and had worked with.

I recall an incident at a workshop held in Bologna, Italy at the 10th International Conference on AI and Law, held at Alma Mater Studiorum – Università di Bologna, in 2005. A student gave a presentation to the workshop which included a discussion of the model of dialogues in Walton and Krabbe [62]. Unknown to the presenter, Doug was sitting in the front row of the audience. Someone asked a question about the dialogue model, to which the presenter responded with a statement that he did not know the answer, and that only the authors of the book would know the answer. Members of the audience who knew Doug laughed, and someone said, "Well, let's ask the author himself!" Doug responded quite humbly, and to the great surprise of the presenter.

References

[1] Model Artificial Intelligence Governance Framework: Second Edition. Personal Data Protection Commission, Government of Singapore, January 21st, 2020.

[2] R. Alur, T.A. Henzinger, and O. Kupferman. Alternating-time temporal logic.

[3] Julia Angwin and Jeff Larson. Bias in criminal risk scores is mathematically inevitable, researchers say. `https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say`, December 30th 2016.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, May 23rd 2016.

[5] Charles Antaki and Ivan Leudar. Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2):181–194, 1992.

[6] Katie Atkinson and Trevor Bench-Capon. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence*, 171(10–15):855–874, 2007.

[37] The main outputs of the Pitlochry symposium are published in [41].

[7] Katie Atkinson, Trevor Bench-Capon, Floris Bex, Thomas F. Gordon, Henry Prakken, Giovanni Sartor, and Bart Verheij. In memoriam Douglas N. Walton: The influence of Doug Walton on AI and law. *Artificial Intelligence and Law*, 28(3):281–326, 2020.

[8] Katie Atkinson, Trevor Bench-Capon, and Peter McBurney. Computational representation of practical argument. *Synthese*, 152(2):157–206, 2006. Section on Knowledge, Rationality and Action.

[9] Katie Atkinson, Rod Girle, Peter McBurney, and Simon Parsons. Command dialogues. In Iyad Rahwan and Pavlos Moraitis, editors, *Proceedings of the Fifth International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2008)*, Lisbon, Portugal, 2008. AAMAS 2008.

[10] Tennent H. Bagley. *Spy Wars*. Yale University Press, New Haven, CT, USA, 2007.

[11] Christian Paz. All the President's Lies About the Coronavirus: An unfinished compendium of Trump's overwhelming dishonesty during a national emergency. The Atlantic, `https://www.theatlantic.com/politics/archive/2020/11/trumps-lies-about-coronavirus/608647/`, November 2nd, 2020. Accessed: 2020-11-14.

[12] Eva Cogan, Simon Parsons, and Peter McBurney. New types of inter-agent dialogues. In Simon Parsons, Nicolas Maudet, Pavlos Moraitis, and Iyad Rahwan, editors, *Argumentation in Multi-Agent Systems*, Lecture Notes in Artificial Intelligence 4049, pages 154–168. Springer, 2006.

[13] B. Jack Copeland. Notes toward a history of possible worlds semantics. In *The Goldblatt Variations: Eight Papers in Honour of Rob*, Uppsala Prints and Preprints in Philosophy, pages 1–14. Department of Philosophy, Uppsala University, Uppsala, Sweden, 1999.

[14] FIPA. Dutch Auction Interaction Protocol Specification. Technical Report XC00032F, Foundation for Intelligent Physical Agents, 15 August 2001.

[15] FIPA. Communicative Act Library Specification. Standard SC00037J, Foundation for Intelligent Physical Agents, 3 December 2002.

[16] Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

[17] Dov M. Gabbay and John Woods. More on non-cooperation in Dialogue Logic. *Logic Journal of the IGPL*, 9(2):321–339, 2001.

[18] Dov M. Gabbay and John Woods. Non-cooperation in Dialogue Logic. *Synthese*, 127(1-2):161–186, 2001.

[19] Jeremy Green. *The Social Construction of the XYY Syndrome*. PhD thesis, University of Manchester, Manchester, UK, 1983.

[20] Jürgen Habermas. *The Theory of Communicative Action: Volume 1: Reason and the Rationalization of Society*. Heinemann, London, UK, 1984. Translation by T. McCarthy of: *Theorie des Kommunikativen Handelns, Band I, Handlungsrationalitat und gesellschaftliche Rationalisierung*. Suhrkamp, Frankfurt, Germany. 1981.

[21] Charles L. Hamblin. *Language and the Theory of Information*. PhD thesis, University of London, London, UK, 1957.

[22] Charles L. Hamblin. Questions in Montague English. *Foundations of Language*,

10(1):41–53, 1973.

[23] Denis J Hilton. Logic and causal attribution. In D. J. Hilton, editor, *Contemporary science and natural explanation: Commonsense conceptions of causality*, pages 33–65. 1988.

[24] Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65–81, 1990.

[25] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the COMPAS recidivism algorithm. `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`, May 23rd 2016.

[26] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[27] Andrea M. Lerner, Gregory K. Folkers, and Anthony S. Fauci. Preventing the Spread of SARS-CoV-2 With Masks and Other "Low-tech" Interventions. *JAMA*, 10 2020.

[28] Stella Lowry and Gordon Macpherson. A blot on the profession. *British Medical Journal*, 296(6623):657–658, 1988.

[29] Igor Lukes. The Rudolf Slánský affair: new evidence. *Slavic Review*, 58(1):160–187, 1999.

[30] Fabrizio Macagno, Douglas Walton, and Chris Reed. Argumentation schemes. History, classifcations, and computational applications. *Journal of Logics and their Applications*, 4(8):2493–2556, 2017.

[31] Jim MacKenzie. What Hamblin's book *Fallacies* was about. *Informal Logic*, 31(4):262–278, 2011.

[32] Robert E. Marks. Validating simulation models: A general framework and four applied examples. *Computational Economics*, 30(3):265–290, 2007.

[33] Peter McBurney, David Hitchcock, and Simon Parsons. The eightfold way of deliberation dialogue. *International Journal of Intelligent Systems*, 22(1):95–132, 2007.

[34] Peter McBurney and Simon Parsons. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language and Information*, 11(3):315–334, 2002. Special Issue on Logic and Games.

[35] Peter McBurney and Simon Parsons. Dialogue games for agent argumentation. In Iyad Rahwan and Guillermo Simari, editors, *Argumentation in Artificial Intelligence*, pages 261–280. Springer, Berlin, Germany, 2009.

[36] Peter McBurney and Simon Parsons. Talking about doing. In Katie Atkinson, Henry Prakken, and Adam Wyner, editors, *From Knowledge Representation to Argumentation in AI, Law and Policy Making: A Festschrift in Honour of Trevor Bench-Capon on the Occasion of his 60th Birthday*, pages 151–166. College Publications, London, UK, 2013.

[37] Jeremy McKean, Hayden Shorter, Michael Luck, Peter McBurney, and Steven Willmott. Technology diffusion: analysing the diffusion of agent technologies. *Autonomous Agents and Multi-Agent Systems*, 17(3):372–396, 2008.

[38] David F. Midgley, Robert E. Marks, and D. Kunchamwar. The building and assurance of agent-based models: An example and challenge to the field. *Journal of Business*

Research, 60(8):884–893, 2007. Special Issue on Complexities in Markets.

[39] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[40] Timothy Miller and Peter McBurney. Efficient storage and retrieval in agent protocol libraries using subsumption hierarchies. *Multiagent and Grid Systems*, 9(2):101–134, 2013.

[41] Chris Reed and Tim Norman (Editors). *Argumentation Machines: New Frontiers in Argument and Computation*. Kluwer Academic, Dordrecht, The Netherlands, 2003.

[42] William Rehg. Reason and rhetoric in Habermas's Theory of Argumentation. In W. Jost and M. J. Hyde, editors, *Rhetoric and Hermeneutics in Our Time: A Reader*, pages 358–377. Yale University Press, New Haven, CN, USA, 1997.

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 1527–1535, 2018.

[44] Thomas Rid. *Active Measures: The Secret History of Disinformation and Political Warfare*. Macmillan, USA, 2020.

[45] Floris Roelofsen. Semantic theories of questions. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, 2019.

[46] Oliver Sacks. *The Man who Mistook his Wife for a Hat*. Summit Books, New York, NY, USA, 1985.

[47] Stefan Şarkadi. *Deception*. PhD thesis, King's College, University of London, London, UK, 2020.

[48] Isabel Sassoon, Sebastian Zillessen, Jeroen Keppens, and Peter McBurney. A formalisation and prototype implementation of argumentation for statistical model selection. *Argument and Computation*, 10(1):83–103, 2019.

[49] Carl E. Schneider. *The Practice of Autonomy: Patients, Doctors and Medical Decisions*. Oxford University Press, Oxford, UK, 1998.

[50] Stewart Steven. *Operation Splinter Factor*. Granada, London, UK, 1974.

[51] Jan Svartvik and Randolph Quirk. *A corpus of English conversation*. Gleerup, Lund, Sweden, 1980.

[52] Douglas Walton. *Arguer's Position: A Pragmatic Study of Ad Hominem Attack, Criticism, Refutation, and Fallacy*. Greenwood Press, Westport, CT, USA, 1985.

[53] Douglas Walton. *Slippery Slope Arguments*. Clarendon Press, Oxford, UK, 1992.

[54] Douglas Walton. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1996.

[55] Douglas Walton. *Appeal to Expert Opinion: Arguments from Authority*. Pennsylvania State University Press, University Park, PA, USA, 1997.

[56] Douglas Walton. A new dialectical theory of explanation. *Philosophical Explorations*, 7(1):71–89, 2004.

[57] Douglas Walton. Examination dialogue: An argumentation framework for critically

questioning an expert opinion. *Journal of Pragmatics*, 38(5):745–777, 2006.

[58] Douglas Walton. Dialogical models of explanation. In *Proceedings of the International Explanation Aware Computing (ExaCt) workshop*, pages 1–9, 2007.

[59] Douglas Walton. A dialogue system specification for explanation. *Synthese*, 182(3):349–374, 2011.

[60] Douglas Walton and Thomas Gordon. Modeling critical questions as additional premises. In F. Zenker, editor, *Argument Cultures: Proceedings of the 8th International Conference of the Ontario Society for the Study of Argumentation (OSSA), May 18-21, 2011.*, pages 1–13, Windsor, ON, Canada, 2011. OSSA.

[61] Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, Cambridge, UK, 2008.

[62] Douglas N. Walton and Erik C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Series in Logic and Language. State University of New York Press, Albany, NY, USA, 1995.

[63] In 1,323 days, President Trump has made 22,510 false or misleading claims. The Washington Post, `https://www.washingtonpost.com/graphics/politics/trump-claims-database/`, September 3rd, 2020. Accessed: 2020-11-14.

[64] Anne L Washington. How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colorado Technology Law Journal*, 17:131, 2018.

[65] Charles Willard. Authority. *Informal Logic*, 12(1):11–22, 1990.

[66] Michael J. Wooldridge. Semantic issues in the verification of agent communication languages. *Journal of Autonomous Agents and Multi-Agent Systems*, 3(1):9–31, 2000.

Received \jreceived