The Bayesian Approach to Robust Argumentation Machines

1. Starting Point

1.1 State of the art and preliminary work

Argumentation is central to our complex world. It pervades law, politics, academia, and everyday life, so it is not surprising that it is the concern of a wide range of disciplines, spanning philosophy, psychology, education, logic and computer science. Philosophers have focussed on normative theories, that is, theories of how we *should* behave. The traditional standard, here, has been formal logic, but more recently, pragma-dialectical theories have focussed on the norms and conventions governing argumentative process (e.g., van Eemeren & Grootendorst, 1992, 2004; Walton, 1995, 1998). Within psychology, 'persuasion' has been a central topic of social psychological research (Petty & Cacioppo, 1986; Eagly & Chaiken, 1993). This has led to a vast literature that has identified many of the moderating variables (e.g., speaker likeability, engagement, mode of presentation, fit with prior beliefs) that affect the degree to which persuasive communication will be effective. Developmental and education research have focussed on the way children's argumentation skills develop and examined ways in which critical thinking and argument skills might be fostered (e.g., Felton & Kuhn, 2001; Kuhn & Udell, 2003; von Aufschnaiter et al., 2008). Logicians and computer scientists have sought to devise argumentation frameworks for dealing with dialectical information, seeking to capture the structural relationships between theses, rebuttals, and supporting arguments with the degree of explicitness necessary for the design of computational argumentation system (Dung, 1995; Prakken & Vreeswijk, 2002; Rahwan & Moraites, 2009).

A shared, focal concern for all of these areas is the issue of *argument quality*: what makes a good argument, and how can good arguments be distinguished from bad ones? This question has two aspects—one descriptive and one normative. On a descriptive level, this question is about success, that is, about what (descriptively) 'works' in convincing others of a position. At the same time, however, researchers in all of the above areas are necessarily engaged in the question of what *should* convince us, and which are the appropriate normative standards against which argument quality should be judged. For philosophers, this question is both part of the longstanding interest in human rationality and part of epistemological concerns about how we can come to secure knowledge of the world (Rescher, 1977; Eva & Hartmann, 2018; Godden & Zenker, 2018). In psychology, the question of argument quality arises for cognitive psychologists descriptively seeking to assess the quality of people's argumentation (e.g., Kuhn, 1991; Corner, Hahn & Oaksford, 2011; Harris et al. 2016) as part of the long tradition of rationality-focussed research on reasoning, judgment and decision-making (Kahneman, 2011). It arises as a target for educational psychologists who want to improve argument skills (von Aufschnaiter et al., 2008), and it arises as a methodological need for social psychologists interested in factors and processes of persuasion that attach to aspects *other* than the intrinsic quality of the argument (Petty & Cacioppo, 1986), because they must experimentally manipulate or statistically *factor out* argument quality. For computer scientists, finally, argument quality matters, because artificial intelligence systems that seek to provide support in complex, high-level domains are ultimately aimed at providing high-quality advice (Jackson, 1998; Neapolitan, 2011), which implies the need to filter, weight and integrate information. So how can argument quality be measured? What normative standards might be devised?

Fallacies of argumentation have historically had a central role in attempts to understand argument quality (Hamblin, 1970; Woods et al., 2004). The longstanding goal of fallacies research was to provide a comprehensive treatment of the catalogue of fallacies that explains exactly why they are 'bad' arguments, ideally giving a unifying *formal account* of why they are poor (Hamblin, 1970). In the last 15 years, the probability calculus (or, where decisions and utilities are involved, Bayesian decision theory) has been applied to the catalogue of fallacies: *arguments from ignorance* (Oaksford & Hahn, 2004; Hahn & Oaksford, 2007, 2008; Harris et al. 2013), *circular arguments* (Hahn & Oaksford, 2007; Hahn, 2011), *slippery slope arguments* (Corner, Hahn & Oaksford, 2011), *ad hominem arguments* (Oaksford & Hahn, 2013), *appeals to expertise* (Harris et al. 2016), *ad populum arguments* (Hahn & Hornikx, 2016), and shorter sketches of how the account applies to the *remainder of the catalogue* of fallacies (Hahn & Oaksford, 2006). This work has shown how a probabilistic treatment can provide a detailed, normative account of argument strength for these fallacies, thus answering long-standing philosophical questions. Because instances of these fallacies occur regularly in everyday discourse (Tindale, 2007), and in some cases even occur frequently (e.g., *ad hominem* arguments), this formal treatment of the fallacies considerably extended the range of normative theories of argument quality. It is also worth noting that this new normative standard did not just address a long-standing theoretical, philosophical concern, it also opened up novel paths for empirical, behavioral experimentation. For each of these fallacies, formalization prompted new experimental programs examining lay people's reasoning (e.g., Corner et al., 2011; Bhatia & Oaksford, 2015; Oaksford & Hahn, 2004; Corner & Hahn, 2009; Harris, Hsu & Madsen, 2012; Hornikx, Harris & Boekema, 2018). Such experiments examine the extent to which arguers intuitively respect normative standards.

More recently, the application of the Bayesian framework to a broader set of schemes for every day argument from the informal logic literature has taken this a step further. The basic idea is easy: One considers an agent (= agent 1) who has (partial) beliefs about a set of propositions and represents these beliefs by a probability function P. Then another agent (= agent 2) wants to convince agent 1 of one of the propositions (= the conclusion). To do so, she proceeds indirectly and endorses some of the other propositions (= the premises) which, in turn, prompts agent 1 to change (typically increase) the probability of these premises. To make sure that the resulting new probability distribution Q is coherent, she updates all other propositions on this new information. In the simplest case, agent 1 considers the premises to be certain after agent 2 endorsed them. The new probability of the conclusion then follows by conditionalizing on these premises. This procedure (slogan: "argumentation is learning") leads to plausible results and can be generalized in various directions, e.g. to cases where the new probability of the premises is not 1, an indicative conditional has exceptions, the sources are only partially reliable, etc. See e.g. Eva & Hartmann (2019), Eva, Hartmann & Singmann (2019), Stern & Hartmann (2018).

The literature on informal argument has catalogued 60+ schemes (Walton et al., 2008; Wachsmuth et al. 2017). Like the fallacies, these schemes identify recurring structures that occur with varying content in everyday discourse; but unlike the fallacies, these particular schemes are taken to be 'good'. Specifically, they provide reasonable, albeit defeasible, inferences for uncertain, ampliative reasoning (which sets them apart from logical schemes such as the classic set of syllogism or conditional reasoning schemes like *modus ponens, modus tollens* etc.). Hahn and Hornikx (2016) showed for several sample schemes how a probabilistic treatment may help achieve the long-held goals of the scheme-based tradition. Here, the Bayesian framework naturally imposes informational relevance through the likelihood ratio. This helps not just with the fallacies, given that fallacies are typically fallacies of relevance (Walton, 1995; 2004), it is essential to capturing argument quality in general. Elucidating in a formally satisfactory, non-question begging way, the notion of relevance itself has been a long-standing challenge (Wilson & Sperber, 1986; Hahn & Oaksford, 2006), and the Bayesian framework helps meet this challenge.

Crucially, the probabilistic treatment of everyday argument schemes provides a solid normative foundation for these schemes that moves beyond mere intuition. This follows from the normative foundation of the Bayesian framework. Here we not only find the traditional pragmatic arguments (a.k.a. Dutch book arguments; Pettigrew, 2020), but also epistemic arguments that show that agents who represent their (partial) beliefs with a probability function minimize the inaccuracy of their (partial) beliefs (Pettigrew, 2016; see also more generally, Corner & Hahn, 2013). Besides its normative justification, the Bayesian framework has the advantage that it connects well with experimental studies from the psychology of reasoning as probability judgments can be extracted from them. While this has been demonstrated for many different reasoning tasks (see, e.g. Oaksford & Chater, 2007), the normative, empirical and computational study of everyday argument schemes in the Bayesian framework is only at the beginning. However, the broad shape of the explanatory project is discernible.

From a philosophical perspective, the Bayesian treatment of fallacies and informal argument schemes brought together formal epistemology and a hitherto separate community concerned with "informal logic". As a result, *Bayesian Argumentation* is now also being expanded to other features of argument (e.g., Zenker, 2014; Godden & Zenker, 2016). At the same time, work by the applicant and colleagues has extended the formal arsenal of Bayesian Argumentation in order to broaden the scope of possible inferences (Eva & Hartmann, 2018b) and has provided detailed treatments of scientific inference schemes (e.g., Dawid, Hartmann & Sprenger, 2015; Eva & Hartmann, 2018; Dardashti & Hartmann, 2019; Dardashti et al. 2019) in a program paralleling the treatment of everyday schemes. Specifically, the applicant and colleagues have shown how new forms of 'evidence' which are not amenable to Bayesian conditionalization (such as indicative conditionals and causal structure) may be captured by minimizing a probabilistic divergence measure from the class of *f*-divergencies (such as the Kullback-Leibler divergence) between the new probability distribution *Q* and the old probability distribution *P*, taking the learned evidence as a probabilistic constraint on *Q* into account (Eva & Hartmann, 2018; Eva, Hartmann & Rafiee Rad, 2020; Stern & Hartmann, 2018).

The body of work on Bayesian Argumentation arguably represents the state of the art with respect to measuring argument quality, in that both a quantitative measure and a well-developed normative basis is provided (see also Nussbaum, 2011). Use of other tools described above frequently shows their limitations: assessment of argument quality within psychology either made do with purely qualitative, structural considerations such as the Toulmin model (Toulmin, 1958; for a critique see Hahn, Zenker & Blum, 2015), relied on pre-testing of intuitively varying materials (for a fundamental methodological critique of this latter strategy see, O'Keefe, 2006), or employed everyday informal schemes whose normative basis still remained somewhat unclear (see Hahn & Hornikx, 2016). Bayesian argumentation has moved argumentation research beyond these limitations.

However, the majority of the work on Bayesian argumentation to date has focussed on single premise-claim relationships. Naturally occurring arguments typically span multiple claims, sub-claims, supporting statements, and complex attack and defeat relations. It is here that qualitative approaches to argument (e.g., Dung et al., 1995; Gordon et al., 2008) in the computational literature have excelled (see e.g., Rahwan & Moraites, 2007). Here too, the need for argument weighting has motivated attempts to add probabilities to these frameworks (e.g., Li, Oren & Norman, 2011), but there are arguably more developed, better suited frameworks for such amalgamation.

With Bayesian Belief Networks (BBNs), the Bayesian framework possesses a computational tool for capturing complex multi-variable relationships. BBNs have been developed in the 1980s as an efficient tool to represent joint probability distributions over a potentially large number of variables (Pearl, 1988; Neapolitan, 2012). This can be achieved by exploiting conditional probabilistic independencies which hold between the variables. These probabilistic independencies form the structure of the BBN which respects the Markov condition ("every variable is independent of its non-descendents, given its parents"); all other conditional independencies can be read-off from the BBN using the *d*-separation criterion. The computational machinery that comes with BBNs is very powerful, which led to successful practical applications in various contexts ranging from defense and military (Falzon, 2006; Laskey and Mahoney, 1997) and cyber security (Chockalingam et al., 2017; Xie et al., 2010), over medicine (Agrahari et al., 2018; Wiegerinck et al., 2013), and law and forensics (Lagnado et al., 2013; Fenton et al., 2013), to agriculture (Drury et al., 2017). On the more theoretical side, BBNs proved to be indispensable tools in computer science (Russell & Norvig, 2020), cognitive science (Gopnik & Tenenbaum, 2007) and parts of philosophy (Bovens & Hartmann, 2003; Sprenger & Hartmann, 2019). Finally, BBNs are closely related to causal graphs which play an important role in the literature on causal discovery (Pearl, 2009; Peters et al., 2017; Sprites et al., 2000). See also Eva, Stern & Hartmann (2019).

Clearly, there are many contexts in which suitable numbers for parameterization of the prior probability distribution (see e.g., Sprenger & Hartmann, 2019, for a general discussion of the problem of priors in the Bayesian framework) may be difficult or impossible to come by, thus necessitating the need for alternative qualitative approaches. However, the fact that BBNs have been used successfully both in practical, AI decision-support systems and in the reconstruction of very large bodies of forensic evidence (Kadane & Schum, 2011), suggests that there should be real-world application domains where BBNs can be deployed successfully for argumentation.

Moreover, there is a strand of research originating in the late 1990's that has examined automatic argument generation from BBNs (Zukerman et al., 1998; Zukerman et al., 1999). This line of research has recently converged with the considerable interest in explainable AI and the automatic generation of explanations within BBNs (Gunning and Aha, 2019; Tešić and Hahn, forthcoming)—a development that is made plausible by the close connections between argument and explanation both conceptually (Bovens and Hartmann, 2003; Dardashti et al., 2019; Dizadji-Bahmani et al., 2011; Howson and Urbach, 2006; Tešić, 2019) and psychologically (Hahn and Oaksford, 2006; Hahn and Oaksford, 2007; Hahn and Hornikx, 2016; Rehder, 2014; Rottman and Hastie, 2014; Tešić and Hahn, 2019; Tešić et al., 2020).

For BBNs to support robust argumentation machines in real-world practice, however, more is required than just suitable algorithms for argument generation and evaluation. Specifically, suitable BBNs must also be learnable from real world data. Here, decades of research on learning BBNs (Geiger and Heckerman, 1994; Heckerman et al., 1999; Chickering, 2002; Neapolitan, 2003; Pearl, 2009; Peters et al., 2017; Spirtes et al., 2000) can be brought to bear. At the same time, empirical evaluation has been shown to be crucial in assessing AI systems when the output consists of outputs that are supposed to be understandable to a human user, as, e.g., in explanation providing recommender systems, Zhang & Chen, 2020). Given the close conceptual connections between explanation and argument, we think that such empirical exploration should take place in assessment of argumentation machines.

In this proposal, we seek to bring together these currently disparate strands of research to provide a) automatic argument generation and evaluation across b) structured many-variable domains including multiple, interacting premises and conclusions, based on c) machine learning of BBN's from large real world data sets, which we will put to d) rigorous evaluation by human users.

2. Objectives and work program

2.1 Objectives

It is well-known that the Bayesian approach to argumentation (i) has a solid normative foundation and (ii) connects well with empirical data from experiments in the psychology of reasoning and argumentation. The main objective of this research proposal is to demonstrate that it also has the computational resources to allow for large-scale applications in the context of robust argumentation machines. We will adapt some of the available computational tools and methods to the study of argumentation and develop new tools and methods if needed. More specifically, our project has the following four objectives:

1. To use machine learning tools to learn BBNs from large data sets.
2. To develop adequate argument generation and evaluation algorithms from these BBNs.
3. To set up tools for testing perceived argument quality of generated arguments.
4. To use these tools to test the arguments we generated.

2.2 Work program incl. proposed research methods

We divide the proposed research into four interrelated work packages (WPs).

WP1: Learning BBNs from data sets

The methods for learning the structure of BBNs can be grouped in two categories (Peters et al., 2017): (i) Independence-based methods relate the conditional independence in data with *d*-separations in the graph. Algorithms such as the IC algorithm (Pearl, 2009), the SGS algorithm, and the PC algorithm (Spirtes et al., 2000) use conditional probability tests to answer queries regarding *d*-separations in the graph. They first build undirected graphs that underlie the network structure (so-called 'skeletons'). Then the directionality is set by: (a) finding v-structures whereby two mutually disconnected variables (e.g. A and B) are connected to a third variable (e.g. C) and A and B are not conditionally independent given C; in these cases a directed arrow is drawn from A to C and from B to C; and (b) respecting the acyclicity constraint, i.e. that a BBN cannot contain cycles.

(ii) Score-based methods, on the other hand, score each (initially randomly generated) candidate BBN's fit to data and then search for ways to modify these networks in order to maximize the score (Geiger and Heckerman, 1994, Heckerman et al., 1999, Chickering, 2002). Modifications are often done using a greedy algorithm (e.g. hill-climbing or tabu search) where a pair of variables is picked and one of three actions is performed and their score assessed: (a) add an arrow, (b) delete an existing arrow, or (c) change the direction of the arrow. The action with the highest score will be performed and another pair of the variables will be chosen until the score does no longer significantly improve.

For learning BBNs from data, we will be using a freely available R package 'bnlearn' (https://www.bnlearn.com/). The package implements both independence-based and score-based algorithms for learning BBNs along with a number of score functions and independence tests. We will use both the independence-based and score-based methods to learn BBNs from data. On the independence-based methods side we will use the si.hiton.pc function that implements the Semi-Interleaved HITON-PC algorithm. From the score-based methods we will use tabu function that implements the tabu greedy search algorithm and Bayesian information criterion as a scoring function. 'bnlearn' also allows expert knowledge input in terms of which arrows should be connected and in which direction. Where that knowledge is available we will use it as additional constraints to learning BBNs.

We will use a publicly available data set on financial factors that drive stock returns for learning BBNs (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). This consists of factors constructed based on both the U.S. returns and international returns. The factors and returns will then be used as variables in learning BBN structures. The discussion on what factors drive stock returns is still ongoing in the finance literature (Fama and French, 1993; 2012; 2015), but some of the factors that have been identified as driving the stock returns are: value (a relationship between the stock price and book value of a company), size (market capitalization or total dollar value of all outstanding shares of a company), momentum (past performance of stocks), quality (capures metrics like debt, earnings growth, quality of management, corporate governance, etc.), volatility (fluctuations in stock prices relative to the market). How many and which factors one should include in explaining the stock returns is very much part of the current debate in finance (Blitz et al., 2018). In particular, it is not clear how these factors interact with each other. Learning BBNs on investment factors data is thus a promising starting point. New (big) data are, however, rapidly becoming available in various domains; while our initial cycle will focus on this economic data set, we will repeat the entire process with new data sets for robustness, choosing the best available at that point in time.

WP1 summary: This WP will involve 1) data preparation 2) application of the learning algorithms to yield the deliverable of a BBN representing the domain which will then form the basis for argument generation and evaluation.

WP2: Automated argument generation and evaluation

Some of the recent examples of argument generation from BBNs resulted from the BARD project (Cruz et al., 2020; Dewitt et al., 2018; Liefgreen et al., 2018; Nicholson et al., 2020; Phillips et al., 2018; Pilditch et al., 2018; Pilditch et al., 2019). This project has set as its goal the development of assistive technology that could facilitate group decision-making in an intelligence context. To this end, BARD provided a graphical user interface enabling intelligence analysts to represent arguments as BBNs and allowing them to examine the impact of different pieces of evidence on arguments. An essential component of BARD is an algorithm for generating natural language explanations of inference in a BBN, or more specifically, an explanation of evidence propagation in a BBN. This algorithm builds on earlier work by Zukerman and colleagues that have sought to use BBNs to generate arguments (Zukerman et al., 1998; Zukerman et al., 1999). The algorithm uses an evidence-to-goal approach to generate explanations for a BBN. An explanation starts with the given pieces of evidence and traces paths that describe their influence on intervening nodes until the goal is reached. In essence, the algorithm adopts a causal interpretation of the links between the connected nodes, finds a set of rules that describe causal relations in a BBN, and calculates all paths between evidence nodes and target nodes and builds corresponding trees in order to determine the impact of evidence on target nodes.

Although this and similar algorithms are a big step towards generating arguments and explanations from BBNs, there are at least two problems related to these algorithms: (1) The algorithm retains difficulties in coping adequately with soft evidence. Namely, the current versions of the algorithm are not able to calculate the impact of evidence that has been learnt with probability less than 1. (2) The explanations and arguments generated by the algorithm are not aimed specifically at what a human user might find hard to understand, i.e. the interactions between variables that often have counterintuitive effects (for psychological evidence to this effect see for example Dewitt et al. 2018; Liefgreen et al. 2018; Phillips et al. 2018; Pilditch et al. 2018; Pilditch et al. 2019; Tešić et al. 2020). What is needed is an algorithm that provides explanations that human users can understand. The goal of this work package is to develop algorithms that solve these problems.

To do so, we need a combination of conceptual, empirical and formal approaches. Problem (1) requires the exploration of updating methods that differ from Bayesian conditonalization (“Bayes Rule”). A straightforward generalization of Bayes Rule is Jeffrey conditonalization. However, this rule cannot be used if additional constraints apply and the so-called *rigidity condition* is violated (Talbott 2016). Such constraints (e.g. that certain marginal or conditional probabilities are kept fixed) are quite typical in complex argumentation situations and require a more sophisticated formal treatment. We propose to use the distance-based approach to Bayesianism developed in Eva, Hartmann, and Rafiee Rad (2020).

According to this approach, one considers an agent who has a prior probability distribution P over a set of variables who then learns some new information. This information is translated into a probabilistic constraint on the new distribution *Q*. For example, if the agent learns that E is the case, then *Q*(E) = 1. If the agent learns that the new probability of E is *e*’ < 1, then *Q*(E) = *e*’ < 1.To determine the full new distribution *Q*, the agent then minimizes an *f*-divergence between *Q* and *P*. Interestingly, if the only constraint that applies is *Q*(E) = *e*’, then one recovers Jeffrey Conditionalization for all *f*-divergencies. The power of the approach, however, lies in its potential to deal with more complex evidential situations in which, e.g., the probability values of various other variables are kept fixed or a conditional probability is set to a new value (thus violating the rigidity condition). Using this approach is normatively justified (Eva et al., 2020) and has already been successfully applied to the problem of Bayesian argumentation (Eva and Hartmann, 2018). We therefore expect the approach to successfully address problem (1). However, the *f*-divergencies approach is computationally much more expensive than conditionalization. Hence it is one of the challenges of WP2 to make its algorithmic implementation more efficient.

To address problem (2), we will empirically evaluate the outputs of existing algorithms which will not only give us an insight into how human users perceive these outputs, but also inform the development of new versions of argument generation algorithms. Further inspiration for the development of new algorithms will come from the literature in cognitive science and in the philosophy of science on what a good explanation is. Several decades of research in these fields have brought on significant insights on what would or would not make an explanation a good one. Most relevant in this context is the work by Tanya Lombrozo and collaborators (cognitive science), e.g. Lombrozo and Vasilyeva (2017), and James Woodward (philosophy), esp. Woodward (2003). To arrive at better algorithms, we expect a back-and-forth between new proposals (informed by empirical and conceptual considerations) and empirical studies, in a cyclical process of development. In short, WP2 represents the conceptual heart of the whole project and will, in one form or other, be a focal point throughout the duration of the project.

Summary WP2: This work package will generate as deliverables a set of metrics and algorithms for argument generation and evaluation, which can be compared with human judgments of argument quality.

WP3: Empirical (user) judgment of arguments: Building the tools

We will build an integrated web environment for testing both individual arguments, and for examining the success of arguments in debate, thus breaking new ground toward the introduction of a range of new ecologically valid measures into argument evaluation.

There is now a sizable body of research spanning more than 15 years which has used scenario-based, behavioral experiments to examine the extent to which lay arguers share the quality considerations explicated by the Bayesian, probabilistic framework (see for a review, Hahn, 2020). In WP3, we will use these methods to examine the extent to which our algorithmic evaluations of argument quality and our algorithmically generated arguments match human intuition.

However, we will also move to break new methodological ground here. Public debate involves arguments and argumentative exchanges that are aimed not just at a specific addressee, but also at a wider audience that is following the argumentative exchange. This prompts the need for theorizing about 'argumentative success' beyond the immediate dialectical exchange of a proponent and opponent. Both argumentation and persuasion research have historically focussed on dyads—communicating pairs exchanging reasons for claims in an ongoing exchange that involves competing and supporting arguments that combine complex, often hierarchically nested ways. Researchers have sought to understand both argument 'quality' and persuasive 'success' within that dyadic frame of reference, developing both procedural rules for engagement and graphing techniques or 'maps' in aid of constructive argument evaluation and production (van Eemeren & Grootendorst, 2003; Gordon, Prakken & Walton, 2007). This scales only partly to contexts with multiple, and possibly large numbers of communicating agents (see also Bonevac, 2003; Lewinski & Aakhus, 2014). *To study this, we will set up a new social media-like forum in which participants are invited to debate specified topics* (see e.g., Becker, Brackbill & Centola, 2017). Debates can then be 'seeded' with particular arguments for evaluation. Specifically, one can track not only how these arguments fare subsequently (are they taken up by others?, are they subject to rebuttal?, how many likes do they receive?). Hence the forum will provide a platform for the collection of important qualitative and quantitative data.

We will combine survey tools for scenario-based analysis and this new forum into a single, integrated environment for argument quality evaluation, thus producing a new, multi-purpose research tool for argumentation research. Technically, this can be achieved by bundling survey software with available toolboxes for interactive comment pages such as the GD bbPress Toolbox Pro for Wordpress.

Summary WP3: This work package will provide as a deliverable an integrated environment for testing human participants' judgments of argument quality, both explicitly and implicitly; this environment will comprise two components, a survey tool for explicit scenario-based evaluation, and a debate forum for implicit assessment of argument quality. This tool will not only support our own subsequent testing but will be made available to other researchers.

WP4: Empirical (user) judgment of arguments: Testing automatically generated argument generation and evaluation

Empirical testing of outputs of AI systems has been a common place in certain domains of AI such as recommender systems for decades now (Zhang and Chen, 2020). In our project we will similarly use the integrated web environment to empirically evaluate the automatically generated arguments from BBNs. More specifically, we will test the generated arguments by analyzing how people using the web environment judge the quality of these arguments. The participants in the experiments will be recruited from online participant recruiting platforms such as Amazon Mechanical Turk and Prolific Academic. They will be provided with monetary compensation.

The results of this empirical evaluation clarify how the general population judges the quality of the generated arguments, thus informing potential improvements of the algorithms for the generations of these arguments. This is important from both a psychological aspect of understanding of argument evaluation and the computer science aspect of algorithm development. Specifically, on the topic of arguments learned from the economic data, we will conduct two scenario-based survey studies and two debate runs. In these debate runs, participants are free to introduce whatever arguments they see fit, but the research assistant will also ‘feed in’ the critical arguments we want evaluated. We will then be able to track how these arguments fared.

Summary WP4: This work package will 1) develop measures for implicit argument quality evaluation in a debate context (including, e.g., “likes”, “uptake”) and 2) conduct 8 behavioral studies (N = 80 participants to achieve sufficient power for statistical analysis). On each topic, there will be two scenario-based studies and two debates.

Research methods. This project will make use of and combine a wide variety of methods, ranging from formal modelling techniques, machine learning and psychological experimentation to conceptual analysis. The success of the project will depend on the skilled combination of these methods. The applicant and the proposed Mercator Fellow have used these methods in previous investigations (e.g. in Collins et al., 2020 and Hahn & Hartmann, 2020) and therefore expect to successfully carry out the proposed research.

3. Bibliography concerning the state of the art, the research objectives, and the work programme

Agrahari, R., Foroushani, A., Docking, T.R., Chang, L., Duns, G., Hudoba, M., Karsan, A., & Zare, H. (2018). Applications of Bayesian network models in predicting types of hematological malignancies. *Scientific Reports, 8*(1), 6951.

Aufschnaiter, C. von, Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching*, *45*(1), 101-131.

Banerjee, A., Gou, X., & Wang, H., (2005). On the optimality of conditional expectation as a Bregman predictor, *IEEE Trans. on Information Theory*, 51 (7).

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 114(26), E5070-E5076.

Bhatia, J.S., & Oaksford, M. (2015). Discounting testimony with the argument ad hominem and a Bayesian congruent prior model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1548.

Blitz, D., Hanauer, M.X., Vidojevic, M., & van Vliet, P. (2018). Five concerns with the five-factor model. *The Journal of Portfolio Management*, *44*(4), 71-78.

Bonevac, D. (2003). Pragma-dialectics and beyond. *Argumentation*, *17*(4), 451-459.

Bovens, L., & Hartmann, S. (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Chickering, D.M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, *3*, 507–554.

Chockalingam, S., Pieters, W., Teixeira, A., & van Gelder, P. (2017). Bayesian network models in cyber security: a systematic review. *Nordic Conference on Secure IT Systems*, 105–122.

Collins, P.J., Krzyzanowska, K., Hartmann, S., Wheeler, G. and Hahn, U. (2020). Conditionals and Testimony. To appear in *Cognitive Psychology*.

Corner, A., & Hahn, U. (2009). Evaluating science arguments: evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, *15*(3), 199.

Corner, A., & Hahn, U. (2013). Normative theories of argumentation: are some norms better than others? *Synthese*, *190*(16), 3579-3610.

Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, *64*(2), 133-152.

Cruz, N., Desai, S.C., Dewitt, S., Hahn, U., Lagnado, D., Liefgreen, A., Phillips, K., Pilditch, T., & Tešić, M. (2020). Widening access to Bayesian problem solving. *Frontiers in Psychology, 11*, 660.

Dardashti, R., & Hartmann, S. (2019). Assessing scientific theories: The Bayesian approach, in: R. Dardashti R. Dawid & K. Thébault (eds.), *Why Trust a Theory?* Cambridge: Cambridge University Press, 67–83.

Dardashti, R., Hartmann, S., Thébault, K., & Winsberg, E. (2019). Hawking radiation and analogue experiments: A Bayesian analysis. *Studies in History and Philosophy of Modern Physics* 67, 1–11.

Dawid, R., Hartmann, S., & Sprenger, J. (2015). The no alternatives argument. *The British Journal for the Philosophy of Science*, *66*(1), 213-234.

Dewitt, S., Lagnado, D.A., & Fenton, N.E. (2018). Updating prior beliefs based on ambiguous evidence. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Drury, B., Valverde-Rebaza, J., Moura, M., & de Andrade Lopes, A. (2017). A survey of the applications of Bayesian networks in agriculture. *Engineering Applications of Artificial Intelligence, 65*, 29–42.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, *77*(2), 321-357.

Eagly, A. H., & Chaiken, S. (1993). *The Psychology of Attitudes*. Harcourt Brace Jovanovich College Publishers.

Eemeren, F. H. van, & Grootendorst, R. (1992). *Argumentation, Communication, and Fallacies*: *A Pragma-dialectical Perspective*. Hilldale, NJ: Lawrence Erlbaum.

Eeemeren, F.H. van, & Grootendorst, R. (2004). *A Systematic Theory of Argumentation. The Pragma-Dialectical Approach.* Cambridge: Cambridge University Press.

Eva, B., & Hartmann, S. (2018). When no reason for is a reason against. *Analysis*, *78*(3), 426-431.

Eva, B., & Hartmann, S. (2018b). Bayesian argumentation and the value of logical validity. *Psychological Review,* 125(5), 806-821.

Eva, B., Stern, R., & Hartmann, S. (2019). The similarity of causal structure. *Philosophy of Science* 86(5), 821-835.

Eva, B., Hartmann, S., & Rafiee Rad, S. (2020). Learning from conditionals. *Mind* 129(514), 461-508.

Falzon, L. (2006). Using Bayesian network analysis to support centre of gravity analysis in military planning. *European Journal of Operational Research, 170*(2), 629–643.

Fama, E.F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33,* 3-56.

Fama, E. F., & French, K. R. (2012). Size, value, and momentum in international stock returns. *Journal of Financial Economics*, *105*(3), 457-472.

Fama, E.F., & French, K.R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1-22.

Felton, M., & Kuhn, D. (2001). The development of argumentative discourse skill. *Discourse Processes*, *32*(2-3), 135-153.

Fenton, N., Neil, M., & Lagnado, D.A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science, 37*(1), 61–102.

Fox, J., Glasspool, D., & Bury, J. (2001). *Quantitative and Qualitative Approaches to Reasoning Under Uncertainty in Medical Decision Making.* Heidelberg: Springer.

Geiger, D., & Heckerman, D. (1994). *Learning Gaussian Networks*. Technical report, Microsoft Research, Redmond, Washington. Available as Technical Report MSR-TR-94-10.

Gelder, T. van (2002). Argument mapping with reasonable. *The American Philosophical Association Newsletter on Philosophy and Computers*, *2*(1), 85-90.

Godden, D., & Zenker, F. (2018). A probabilistic analysis of argument cogency. *Synthese*, *195*(4), 1715-1740.

Gopnik A., & Tenenbaum J.B. (2007). Bayesian networks, Bayesian learning, and cognitive development. *Developmental Science* 10(3), 281-287.

Gordon, T. F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, *171*(10-15), 875-896.

Gunning, D., & Aha, D.W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine, 40*(2), 44–58.

Hahn, U. (2011). The problem of circularity in evidence, argument and explanation. *Perspectives on Psychological Science,* 6, 172-182.

Hahn, U. (2020). Argument quality in real world argumentation. *Trends in Cognitive Sciences*, 24(5), 363-374.

Hahn, U., & Hornikx, J. (2016). A normative framework for argument quality: Argumentation schemes with a Bayesian foundation. *Synthese,* 193, 1833-1873.

Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese*, 152, 207-236.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*,114, 704-732.

Hahn, U., & Oaksford, M. (2007b). The burden of proof and its role in argumentation. *Argumentation*, 21, 39-61.

Hahn, U., & Oaksford, M. (2008). Inference from absence in language and thought. In: N. Chater and M. Oaksford (eds.). *The Probabilistic Mind*, Oxford University Press, 121-142.

Hahn, U., Blum, R., & Zenker, F. (2017). Causal argument. M. Waldmann (ed.): *The Oxford Handbook of Causal Reasoning.* Oxford: Oxford University Press.

Hahn, U., & Hartmann, S. (2020). A new approach to testimonial conditionals. To appear in *CogSci 2020*.

Hamblin, C.L. (1970). *Fallacies.* London: Methuen.

Harris, A.J., Hsu, A.S., & Madsen, J.K. (2012). Because Hitler did it! Quantitative tests of Bayesian argumentation using ad hominem. *Thinking & Reasoning*, *18*(3), 311-343.

Harris, A.J.L., Hahn, U., Madsen, J.K., & Hsu, A.S. (2016). The appeal to expert opinion: quantitative support for a Bayesian network approach. *Cognitive Science*, 40*,* 1496-1533.

Heckerman, D., Meek C., & Cooper, G. (1999). A Bayesian approach to causal discovery. In: C. Glymour and G. Cooper (eds.): *Computation, Causation, and Discovery*, 141–165. Cambridge, MA: MIT Press.

Hornikx, J., Harris, A. J., & Boekema, J. (2018). How many laypeople holding a popular opinion are needed to counter an expert opinion? *Thinking & Reasoning*, *24*(1), 117-128.

Jackson, P. (1998). *Introduction to Expert Systems*. Addison-Wesley.

Kadane, J.B., & Schum, D.A. (2011). *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. John Wiley & Sons.

Kahneman, D., & Egan, P. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Korb, K.B., & Nicholson, A.E. (2010). *Bayesian Artificial Intelligence*. CRC Press.

Kuhn, D. (1991). *The Skills of Argument.* Cambridge: Cambridge University Press.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development*, *74*(5), 1245-1260.

Lewiński, M., & Aakhus, M. (2014). Argumentative polylogues in a dialectical framework: A methodological inquiry. *Argumentation*, *28*(2), 161-185.

Lagnado, D.A., Fenton, N., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation, 4*(1), 46–63.

Li, H., Oren, N., & Norman, T.J. (2011). Probabilistic argumentation frameworks. In: *International Workshop on Theory and Applications of Formal Argumentation*. Berlin, Heidelberg: Springer, 1-16.

Laskey, K.B. & Mahoney, S. M. (1997). Network fragments: Representing knowledge for constructing probabilistic models. *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, 334–341, Morgan Kaufmann Publishers.

Liefgreen, A., Tešić, M., & Lagnado, D. (2018). Explaining away: significance of priors, diagnostic reasoning, and structural complexity. *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*

Lippi, M., & Torroni, P. (2015). Argument mining: A machine learning perspective. In *International Workshop on Theory and Applications of Formal Argumentation*. Springer*,* 163-176.

Lombrozo T., & Vasilyeva N. (2017). Causal explanation. In M. Waldmann (ed.): *Oxford Handbook of Causal Reasoning*. Oxford UK: Oxford University Press, 415-432.

Neapolitan, R.E. (2003). *Learning Bayesian Networks*. Upper Saddle River, NJ: Pearson Prentice Hall.

Neapolitan, R.E. (2012). *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. Create Space Independent Publishing Platform.

Nicholson, A.E., Korb, K. B., Nyberg, E. P., Wybrow, M., Zukerman, I., Mascaro, S., Thakur, S., Alvandi, A.O., Riley, J., Pearson, R. et al. (2020). BARD: A structured technique for group elicitation of bayesian networks to support analytic reasoning. *arXiv:2003.01207*.

Nussbaum, E.M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist*, *46*(2), 84-106.

Oaksford, M., & Chapter, N. (2007). *Bayesian Rationality*. Oxford: Oxford University Press.

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, 58, 121-131.

Oaksford, M., & Hahn, U. (2013). Why are we convinced by the ad hominem argument? Source reliability or pragma-dialectics? In: Zenker, F. (ed.): *Bayesian Argumentation*. Springer Library, 39-58.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of Causal Inference*. Cambridge, MA: The MIT Press.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Pettigrew, R. (2020). *Dutch Book Arguments*. Cambridge: Cambridge University Press.

Petty, R. E., & Cacioppo, J.T. (1986). The elaboration likelihood model of persuasion. In: *Communication and Persuasion*. New York: Springer*,* 1-24.

Phillips, K., Hahn, U., & Pilditch, T.D. (2018). Evaluating testimony from multiple witnesses: single cue satisficing or integration? *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Pilditch, T.D., Fenton, N., & Lagnado, D. (2019). The zero-sum fallacy in evidence evaluation. *Psychological Science, 30*(2), 250–260.

Pilditch, T. D., Hahn, U., & Lagnado, D. A. (2018). Integrating dependent evidence: na ̈ive reasoning in the face of complexity. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Prakken, H., & Vreeswijk, G. A.W. (2002). Logics for defeasible argumentation. In D.M. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic, 2nd edition, Vol 4,* 219-318. Dordrecht: Kluwer.

Rahwan, I., & Moraites, P. (eds.) (2009). *Argumentation in Multi-Agent Systems. Fifth International Workshop, ArgMAS 2008. Lecture Notes in Computer Science*, Vol. 5384, Heidelberg: Springer.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72*, 54–107.

Rescher, N. (1977). *Dialectics: A Controversy Oriented Approach to the Theory of Knowledge*. Albany, NY: SUNY Press.

Rottman, B.M., & Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology, 87*, 88–134.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th Edition). Pearson.

Spirtes, P., Glymour, C., & Scheines R. (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Sprenger, J. & Hartmann, S. (2019). *Bayesian Philosophy of Science*. Oxford: Oxford University Press.

Stern, R., & Hartmann, S. (2018). Two sides of modus ponens. *The Journal of Philosophy,* 115 (11), 605–621.

Talbott, W. (2016). Bayesian Epistemology. In: *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian/>.

Tešić, M. (2019). Confirmation and the generalized Nagel–Schaffner model of reduction: a Bayesian analysis. *Synthese, 196*(3), 1097–1129.

Tešić, M., Liefgreen, A., & Lagnado, D. (2020). The propensity interpretation of probability and diagnostic split in explaining away. *Cognitive Psychology, 121*, 101-293.

Tešić, M. & Hahn, U. (2019). Sequential diagnostic reasoning with independent causes. *Proceedings of the 41th Annual Conference of the Cognitive Science Society.*

Tešić, M. & Hahn, U. (forthcoming). Explanation in AI systems. In *Human-like Machine Intelligence.* Oxford University Press*.*

Tindale, C.W. (2007). *Fallacies and Argument Appraisal*. Cambridge: Cambridge University Press.

Toulmin, S.E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., …, & Stein, B. (2017). Computational argumentation quality assessment in natural language. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1*, Long Papers, 176-187.

Walton, D.N. (1995). *A Pragmatic Theory of Fallacy.* Tuscaloosa/London: The University of Alabama Press.

Walton, D.N. (1998). *The New Dialectic: Conversational Contexts of Argument.* Toronto: University of Toronto Press.

Walton, D. (2004). Classification of fallacies of relevance. *Informal Logic*, *24*(1), 73-103.

Walton, D. & Macagno, F. (2016). A Classification System for Argumentation Schemes. *Argument & Computation,* 1-29.

Walton, D.N., Reed, C., & Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press.

Wiegerinck, W., Burgers, W., & Kappen, B. (2013). Bayesian networks, introduction and practical applications. *Handbook on Neural Information Processing*. Springer, 401–431.

Wilson, D., & Sperber, D. (1986). On defining relevance. In: R. Grandy & R. Warner (eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. Oxford: Oxford University Press, 243-258.

Woods, J., Irvine, A., & Walton, D.N. (2004*). Argument: Critical Thinking, Logic and the Fallacies,* Revised Edition. Toronto: Prentice Hall.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Xie, P., Li, J. H, Ou, X., Liu, P., & Levy, R. (2010). Using Bayesian networks for cyber security analysis. *2010 IEEE/IFIP International Conference on Dependable Systems & Networks (DSN)*, 211–220.

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval, 14*(1)*,* 1–101.

Zukerman, I., McConachy, R., & Korb, K.B. (1998). Bayesian reasoning in an abductive mechanism for argument generation and analysis. *AAAI/IAAI*, 833–838.

Zukerman, I., McConachy, R., Korb, K.B., & Pickett, D. (1999). Exploratory interaction with a Bayesian argumentation system. *IJCAI*, 1294–1299.