Final Report:
Evaluating dialectical explanations for recommendations

Partners:
- Francesca Toni (Imperial College London, Computing) – PI
- Dave Lagnado and Christos Bechlivanidis (UCL, Experimental Psychology) – CoIs
- Antonio Rago (Imperial College London, Computing) – PostDoc

The project aimed at addressing the lack of transparency of AI techniques, e.g. machine learning algorithms or recommender systems, one of the most pressing issues in the field, especially given the ever-increasing integration of AI into everyday systems used by experts and non-experts alike, and the need to explain how and/or why these systems compute outputs, for any or for specific inputs. The need for explainability arises for a number of reasons: an expert may require more transparency to justify outputs of an AI system, especially in safety-critical situations, while a non-expert may place more trust in an AI system providing basic (rather than no) explanations, regarding, for example, films suggested by a recommender system. Explainability is also needed to fulfil the requirements of the forthcoming General Data Protection Regulation (GDPR), becoming effective from May 25th, 2018. Indeed GDPR can be interpreted as effectively creating a right-to-explanation for users of automated decision-making systems [1]. Furthermore, explainability is crucial to guarantee comprehensibility in Human-Like Computing, to support collaboration and communication between machines and human beings.

The role of explanation in AI can be seen as threefold [2]: i) to inform why a particular output was produced; ii) to provide means to contest the output if undesired/unexpected and iii) to help understand what could be changed to get the desired/expected output. Various forms of explanation have been proposed. For example, in [2] explanations are seen as counter-factual statements, of the form "score p was returned because variables V had values v; if V instead had values v' then score p' would have been returned"; in [3] explanations are extracted from text in natural language and in [4] from images as evidence for predictions of classifications, obtained from interpretable models learnt locally around the predictions; further, in [5,6] explanations are sets of logical rules, learnt from data, and in [7] they are (parts of) argumentation graphs, as understood in the field of computational argumentation in AI [8].

Computational argumentation is suitable to deal in particular with reasoning and decision-making in contexts where contradictions/inconsistencies/conflicts arise, information on which to base reasoning and decision-making is incomplete, hypotheses need to be made and assessed against one another and in the context where additional, possibly probabilistic, information may or may not be present. Several computational argumentation-based approaches for reasoning and decision-making have been presented, many equipped with explanatory functionalities, in that they can justify their conclusions and/or recommendations by means of dialectical explanations (e.g. see [7,9]) which unveil how the underlying contradictions/inconsistencies/ conflicts have been resolved during reasoning).

There is also a widespread belief in the AI community that humans can relate to dialectical explanations better than to other forms of explanations, somewhat corroborated by indications, in the psychology literature, that humans developed reasoning in order to argue [10]. With very few exceptions though (notably [11-13]) there has been hardly any empirical evaluation with humans as to the usefulness of the outputs of computational argumentation tools and methods in AI. Moreover, no experimentation has been conducted as to whether human feedback on dialectical explanations, for example by engaging dialectically with the explanations, may improve computational argumentation tools and methods in AI so that they compute better outputs (conclusions and/or decisions). This gap becomes more alarming in light of recent psychological experiments showing that human reaction to generic (non-

dialectical) explanations is often unexpected and rarely in accordance with philosophical prescriptions. For example, recent research [14] show that, in certain cases humans prefer more complex explanations, i.e. explanations that refer to multiple even overlapping causes. Similarly, while most philosophers stress the importance of abstraction in explanation [15], people appear to prefer even causally irrelevant details [16]. Explanations rich in descriptive information may assist in visualization and, thus, promote a sense of understanding.

In this context, the main aim of this project was *to conduct experiments to determine whether and which computed dialectical explanations*, extracted from the argumentation graphs of [7] for explaining recommendations, *are useful to humans and whether human feedback can improve the outputs of the recommender system.* The planned experiments were identified as useful to confirm or falsify the hypothesis that *argumentation can serve as a paradigm for human-machine interaction*, in the specific setting of recommender systems and argumentative explanations as in [7].

As a secondary aim, we aimed to use the machine output to improve our formal understanding of the explanatory virtues from a human perspective [17], such as the desirable level of concreteness, the complexity of causal models etc.

Outcomes

The project took place during December 2018-April 2019 and carried out the envisaged workplan in the proposal. We implemented an instance of the recommender approach of [7] assuming, as a starting point, an "item-aspect" graph, where items to be recommended (e.g. films) are linked to aspects (e.g. director, actors, genre). In the implementation, users can give (partial) ratings to items, and an algorithm is used to compute user-tailored predicted ratings for items that users have not rated, based on aspects of the items and actual ratings by other users, with parameters indicating how much the aspects and the opinions of the other users should matter, for each user. Then, these graphs are mapped onto argumentation graphs, where, intuitively and informally, items and aspects may be seen as arguments: if a user (or another similar user) rates an item highly/lowly then this item can be seen as an argument for/against, respectively, the aspects connected with the item and, similarly, if a user rates an aspect highly/lowly then this aspect can be seen as an argument for/against, respectively, the items connected with the aspect. Moreover, if an "item-aspect" graph is viewed from an argumentative perspective, a user's (or similar user's) opinion (rating) on an aspect/item may impact the estimation of the user's opinion (rating) of items/aspects connected with that aspect/item in the absence of actual ratings. This argumentative reading of "item-aspect" graphs from the perspective of individual users facilitates the extraction of explanations for predictions, in the form of sub-graphs of argumentation frameworks/graphs, as illustrated in [7]. These (sub-)graphs can be seen as providing a 'back-end' for a variety of explanations in different formats (e.g. graphical, visual or linguistic) for different contexts and types of users. We experimented with various explanation formats and conducted a number of experiments, in specific movie recommender settings, using different forms of explanations drawn from the (sub-)graphs generated by the method of [7]. The experiments were conducted on Amazon Mechanical Turk and are described in the following paper, submitted to the special issue of the Artificial Intelligence journal on Explainable AI:

- Argumentative Explanations for Interactive Recommendations. Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David Lagnado and Francesca Toni

The work identified various directions for future work, including the need for formally defined protocols of interaction for explanatory dialogues with users, motivating further experiments in collaboration with UCL, described in the following paper submitted to KR 2020:

- Argumentation as a Framework for Interactive Explanations for Recommendations. Antonio Rago, Oana Cocarascu, Christos Bechlivanidis and Francesca Toni.

This paper builds upon some preliminary analysis conducted in the following paper:

- From Formal Argumentation to Conversational Systems. Oana Cocarascu, Antonio Rago & Francesca Toni. Proceedings of the 1st Workshop on Conversational Interaction Systems

Following participation in the Human-Like Computing Machine Intelligence workshop - Cumberland Lodge - 30th June to 3rd July 2019, Rago and Toni also prepared, with collaborators, the following paper under revision for inclusion in the MI-HLC2020 OUP Book

- Mining Property-driven Graphical Explanations for Data-centric AI from Argumentation Frameworks. Oana Cocarascu, Kristijonas Cyras, Antonio Rago, Francesca Toni

References

[1]European Union regulations on algorithmic decision-making and a "right to explanation". Goodman and Flaxman. AI Magazine, 38, no. 3 (2017): 50-57.

[2]Counterfactual Explanations without opening the black box: automated decisions and the GDPR. Wachter, Mittelstadt, Russell. Harvard Journal of Law & Technology. In Press.

[3]Rationalizing Neural Predictions. Lei, Barzilay, Jaakkola. EMNLP 2016.

[4]"Why Should I Trust You?" Explaining the Predictions of Any Classifier. Ribeiro, Singh, Guestrin. KDD 2016

[5]Learning Explanatory Rules from Noisy Data. Evans, Grefenstette. Journal of AI Research 61: 1-64 (2018)

[6]Ultra-strong machine learning - comprehensibility of programs learned with ILP. Muggleton, Schmid, Zeller, Tamaddoni-Nezhad, Besold. Machine Learning, 2018. In Press.

[7]Argumentation-Based Recommendations: Fantastic Explanations and How to Find Them. Rago, Cocarascu, Toni. IJCAI 2018. In Press

[8](Formal) Argumentation Theory. Modgil. The Reasoner, 11:6, 2017.

[9]On Computing Explanations in Argumentation. Fan, Toni. AAAI 2015

[10]Why do humans reason? Arguments for an argumentative theory. Mercier, Sperber. BEHAVIORAL AND BRAIN SCIENCES (2011) 34, 57–111.

[11]Behavioral Experiments for Assessing the Abstract Argumentation Semantics of Reinstatement. Rahwan, Madakkatel, Bonnefon, Awan, Abdallah. Cognitive Science 34(8): 1483-1502 (2010)

[12]Formal Arguments, Preferences, and Natural Language Interfaces to Humans: an Empirical Evaluation. Cerutti, Tintarev, Oren. ECAI 2014: 207-212,

[13]Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. Polberg and Hunter. Int. J. of Approximate Reasoning 93: 487-543 (2018).

[14]Evaluating everyday explanations. Zemla, Sloman, Bechlivanidis, Lagnado. Psychonomic Bulletin & Review, pp 1–13, 2017

[15]An account of scientific explanation. Strevens, Depth. Cambridge, MA;Harvard University Press, 2008.

[16]Concreteness and abstraction in everyday explanation. Bechlivanidis, Lagnado, Zemla, Sloman. Psychonomic Bulletin & Review, pp 1–14, 2017

[17]Inference to the best explanation, coherence and other explanatory virtues. Mackonis. Synthese, 190(6), 975–995, 2013