A Unifying Framework for Managing Conflict-laden Content

Max Rapp

FAU Erlangen/Nürnberg

Human-Computer Interaction (HCI) increasingly takes place on the *behavioural* level: human behaviour is tracked and aggregated on an immense scale and used to elicit desired machine behaviour through training. On the human side such behavior is characterised by its low level of consciousness and reflection, the short-term desires and whims that drive it and the small-level decisions that constitute it. Crucially, users who feel anonymous but are in fact highly observable do not intend their exhibited behaviour to shape their online experience. On the other side, machine behaviour is regarded as desirable if it succeeds in predicting and influencing human behaviour. This is achieved through learning algorithms whose output is highly opaque. The result are systems that - while highly effective at giving users what they "want" - do so by exploiting and feeding the biases of "System 1" [2] creating a feedback loop that manipulates and disempowers users by compromising their self-control and their understanding of the systems with which they interact.

This thesis project adheres to the belief that Artificial Intelligence (AI) should instead put users in the driver seat. Thus AI-systems that interact with humans should be based on principles decided upon by users' through a conscious, reflected, high-level process. The inner workings of such systems should be transparent and the reasons for their decisions and actions should be explainable to the users.

The way humans deliberate on and justify their actions or rationalise their behaviour is through arguments. Machines that achieve the goals stated above will therefore require argumentation capabilities: they need to engage user's conscious focus on the options at hand through offering arguments as entry points into reflective processes; they need to assist users in this process through argumentative support and elicit the users' stance on the high-level principles that should guide the HCI; finally, they need to motivate and rationalise their actions through explanations that weigh the arguments in favor of and against each option giving a veracious yet human-understandable representation of the systems actual reasoning process.

The content that needs to be represented, processed and created in these argumentative interactions is highly conflict-laden: not only do arguments frequently support contradicting conclusions (rebuttal), they may also attack other arguments' premises (undercut), the mode of inference they employ or even meta-content such as their utterers. A plethora of argumentation theories has been devised to represent and reason with such content. They comprise frameworks of different levels of formality, intertheoretical integration and implementation. Likewise, the range and depth of application domains for these frameworks is growing rapidly, confronting theory with new and diverse requirements.

This thesis project seeks to develop a unifying knowledge representation framework for argumentation that enables the rapid implementation of use-case adapted formalisms; creating, hosting, sharing, processing and visualizing conflict-laden content across formalisms and application domains; dynamic updating of such content

In: C. Kaliszyk, E. Brady, J. Davenport, W.M. Farmer, A. Kohlhase, M. Kohlhase, D. Müller, K. Pąk, and C. Sacerdoti Coen (eds.): Joint Proceedings of the FMM and LML Workshops, Doctoral Program and Work in Progress at the Conference on Intelligent Computer Mathematics 2019 co-located with the 12th Conference on Intelligent Computer Mathematics (CICM 2019), Prague, Czech Republic, July 8–12, 2019, published at http://ceur-ws.org

through dialogical processes; and ultimately the representation of human agents' minds as such conflict-laden content.

To gather requirements for such a framework and to put its usefulness to the test, a range of applications are intended that roughly correspond to the system features delineated above: the creation of an atlas of argumentation theories; formalization of and reasoning on legal text; a dynamic error and conflict handling system; an argumentation based recommender system for educational content.

As a first step towards these goals the MMT language and system [3] is extended by argumentation capabilities: theory graphs are extended by a defined attack relations yielding what we call *context graphs*. We submit the ALMANAC-hypothesis: "Any argumentation system $\mathcal{A}$ can be refactored into a a classical object-language $\mathcal{L}$ and a context graph scheme $\mathcal{G}$ such that $\mathcal{A}$ is isomorphic to $\langle \mathcal{L}, \mathcal{G} \rangle$." To test the hypothesis, a logic atlas of the existing argumentation theories and their interrelations will be built. MMT's proof checking - and in the future proof assistant capabilities - in combination with an already partially implemented argumentation semantic computation and visualization tool suite for MMT should immediately furnish a prototype implementation for many of the formalisms in the atlas.

The tools developed in the aforementioned extension of MMT are put to use in a project on the formalization of legal text (JLogic). Here proof checking will be employed to assess the correctness of legal arguments found in the text. The created content will be hosted on the MathHub [1] platform.

Later work will see the addition of dynamics to furnish an error and conflict handling system: Automated generation of conflict graphs together with MMT's type checking will yield the capability to provide automated, illuminating error messages and to prompt users on the presence of conflicting content in a theory graph.

Finally, we will explore how far argumentation goes in equipping AI with a theory of mind through a dialogical recommender system for educational content and code documentation: based on students/users queries the system will construct arguments regarding students'/users' current knowledge. Likewise it will attempt to infer the arguments that lead the students'/users' to arrive at erroneous conceptions of the domain. It will use these representations to suggest precisely targeted learning or documentation items to the students'/users.

References

[1] Mihnea Iancu, Constantin Jucovschi, Michael Kohlhase, and Tom Wiesing. System description: Mathhub.info. In Stephen M. Watt, James H. Davenport, Alan P. Sexton, Petr Sojka, and Josef Urban, editors, *Intelligent Computer Mathematics*, pages 431–434, Cham, 2014. Springer International Publishing.

[2] Daniel Kahneman. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York, 2011.

[3] F. Rabe and M. Kohlhase. A Scalable Module System. *Information and Computation*, 230(1):1–54, 2013.