Algorithmic Transparency and Accountability in Practice

Jef Ausloos
Centre for IT and IP Law (CiTiP)
KU Leuven - imec
Sint-Michielsstraat 6 box 3443
3000 Leuven, Belgium
jef.ausloos@kuleuven.be

David Geerts
Meaningful Interactions Lab
KU Leuven - imec
Parkstraat 45 Bus 3605
3000 Leuven, Belgium
david.geerts@kuleuven.be

Pierre Dewitte
Centre for IT and IP Law (CiTiP)
KU Leuven - imec
Sint-Michielsstraat 6 box 3443
3000 Leuven, Belgium
pierre.dewitte@kuleuven.be

Bieke Zaman
Meaningful Interactions Lab
KU Leuven - imec
Parkstraat 45 Bus 3605
3000 Leuven
bieke.zaman@kuleuven.be

Peggy Valcke
Centre for IT and IP Law (CiTiP)
KU Leuven - imec
Sint-Michielsstraat 6 box 3443
3000 Leuven, Belgium
peggy.valcke@kuleuven.be

Abstract
This position paper aims to contribute to the debate on algorithmic transparency and accountability, relating it to compliance with the so-called right to an explanation in EU data protection law. We propose a research agenda based on legal-empirical data, that will constitute the basis for pinpointing key issues, evidence-based policy guidance and conducting further interdisciplinary research. Based on this research agenda, we are preparing the co-creation of a concrete prototype for making recommendation algorithms for news curation understandable to the average individual. This position paper is the result of a collaboration between two research centres, with expertise in law (CiTiP) and Human-Computer Interaction (Mintlab), enabling a more holistic perspective on a critical societal issue.

Author Keywords
Right to explanation; Algorithms; Artificial Intelligence; Data Protection; Human Computer Interaction; Transparency; Accountability.

ACM Classification Keywords
H.5.0. Information interfaces and presentation (e.g., HCI): General;

Introduction

With the rise of Artificial Intelligence (AI) and Machine Learning (ML), technology is increasingly mediating our lives, both online (e.g. media consumption, social networking) and offline (e.g. smart cities). Policy-makers and academia are increasingly looking into the effects of algorithms on society. Despite several initiatives to increase transparency and accountability in this context, there is a manifest lack of empirical research. More specifically, it is not entirely clear how current legal rules are interpreted and applied in practice. European Data Protection law, notably, includes a so-called 'right to an explanation', but there is no academically sound evidence on how it is interpreted and accommodated on the ground. We believe that such trustworthy, empirical data is crucial to better regulation and enforcement, as well as more informed design of user interfaces to accommodate such rights.

In light of this, this position paper proposes a research agenda that contributes to answering one of the key questions in today's information society: how to make algorithmic transparency and accountability work in practice?

The Black Box of Algorithms

Algorithms are everywhere. They increasingly affect what we can find, see and say online, how we interact with our homes and public spaces, our access to jobs, loans, insurance, and justice [13,12]. As such, and for better or worse, algorithms have a growing impact on core democratic values. Problematic in this regard is that currently, much of these algorithms operate behind closed doors. Even if we were to see them, they are opaque, complex and constantly changing [8], making it even harder to get a grasp on them. At the moment, policy-makers and academia are struggling to ensure fundamental norms and values are upheld against this backdrop of exponential economic and technological progress [11]. There are several different ways one can try to break open this expanding black box [15], but one particularly important one can be found in the law. More specifically EU data protection law.

EU Data Protection Law

In May 2018, the General Data Protection Regulation (GDPR) will enter into force. It constitutes a major overhaul of the current regulatory framework, specifically aimed at tackling core challenges faced in the information society. Among the tools to ensure (personal) data is processed fairly, the GDPR provides concrete rights to individuals, ranging from access (Art.15) to erasure (Art.17) and portability (Art.20). One right might prove particularly useful in light of the issues mentioned before: the right to an explanation [10,14]. More specifically, Article 15(1)h empowers individuals to gain access to 'meaningful information about the logic, significance and envisaged consequences of automated decision-making, including profiling'. It can be read together with Article 22, which grants individuals a right to contest such decisions. Controllers are also obliged to implement appropriate technical and organisational measures to comply with these rights and – *de facto* – to ensure effective transparency (Art. 25(1) GDPR).

The exact scope and meaning of the right to explanation in practice has been the subject of vehement debate among scholars in different disciplines [17,10,14]. Indeed, what exactly constitutes 'meaningful information' and how is it influenced by other factors (e.g. age of the individual, context, type of the service)?

Simply showing the underlying algorithms is not a reasonable solution. Apart from the fact that algorithms are impossible to understand as such by humans, there are a number of legitimate counter-arguments that oppose this (e.g. legal defences such as trade secrecy or the constantly changing nature of the algorithm). To what extent can we require individualised explanations (e.g. why this particular recommendation has been made to me), or more generalised explanations (e.g. this is how the algorithm works in broad strokes)? Much of the discussion on this subject, has remained quite theoretical so far. Even if its exact scope can be delineated, the right to explanation is only effective in practice if proper thought is given to how the algorithm is translated and made understandable to the average individual. Enter HCI.

Human Computer Interaction

Human-Computer Interaction (HCI) as a discipline is well suited to study the role of algorithms in our lives, as well as designing new ways of making algorithms and their decisions more understandable and contestable. While there is a clear impact of algorithms on our interaction with technology and other aspects of our lives, such as voting behaviour, many people are not even aware that algorithms are being used to filter information or take important decisions [5]. A first step in designing more understandable algorithmic systems is therefore disclosing if and when an algorithm is being employed [4]. Eslami et al. [5] created FeedVis, a system that showed the difference between the algorithmically curated and an unchanged Facebook News Feed to participants. They found that becoming aware of the algorithm's presence led to more active engagement with the platform and an increased feeling of control. Awareness alone however is not sufficient. Diakopoulos [4] lists four other categories of information that might be disclosed: human involvement (i.e. the goal, purpose, and intent of the algorithm), the data that drives the algorithm (e.g. its accuracy, completeness, uncertainty, timeliness or representativeness), the model itself (including input, weight, training data, assumptions) and inferencing (such as classifications or predictions). Similarly, the ACM U.S. Public Policy Council (USACM) together with the ACM Europe Council Policy Committee (EUACM) call for system designers to ensure fairness by building in various principles into their systems, including "explanation", meaning that the logic of the algorithm, no matter how complex, must be communicable in human terms [7]. Conveying this complexity to users however is not an easy feat. Design experiments with transparency in recommender and other algorithmic systems indicate that providing some explanation increases trust in and acceptation of these systems, but providing too much information reduces trust [9,3]. Moreover, Ananny & Crawford [1] argue that transparency in itself is not sufficient as it has several limitations, and it is crucial to use these limitations as a path towards algorithmic accountability. Further research in designing understandable, transparent and accountable algorithmic systems, based on a thorough understanding of the legal frameworks, is highly needed. This is all the more relevant since ensuring algorithmic transparency and legibility is likely to rank amongst the necessary data protection-by-design measures to be implemented by controllers.

A Research Agenda for Algorithmic Transparency and Accountability

Currently, there is a lack of methodologically sound and objective empirical evidence on how algorithms are explained to individuals in practice. Moreover, the few

academic efforts to develop proof-of-concepts of interfaces to make algorithms understandable generally incorporate little or no legal input. We argue that both of these lacunae need to be tackled by collaboration between the fields of legal and HCI research. Two recent efforts coming from these different perspectives can serve as an example of the challenges and opportunities that such an approach can bring.

In the academic year 2017-2018, the legal research group CiTiP conducted a small-scale pilot, testing a legal-empirical research methodology for gathering data on compliance with data subject rights. Led by a senior researcher, three Master students exercised their rights of access and to erasure with around 60 online service providers. Interactions were systematically mapped in order to enable a critical evaluation of the general state of affairs of these rights *in practice*. Importantly, the research pointed to deeper non-legal issues related to the complexity of algorithms and how to make them understandable to the average individual [2]. This inspired CiTiP to take a more holistic approach and seek active collaboration from other disciplines, in particular the field of HCI.

In the same period, the HCI research group Mintlab concluded a research project where algorithms were being used to support home care planners through automated scheduling techniques and decision support software. While the human task planners were positive with the received support, it was paramount for them to gain an understanding of how the decisions were being made by the system. Researchers from Mintlab wanted to continue this line of research, but needed deeper insight into legal implications, notably regarding the right of access and explanation. Indeed, only through intensive cooperation between researchers in both HCI and law, will it be possible to give meaning to a right that is ultimately aimed at providing more accountability in today's complex information society.

We propose an interdisciplinary collaborative research agenda to explore how key challenges (to individual freedoms, fundamental rights and societal values) emanating from the increased reliance on algorithms, can be thwarted via - so far underused - data subjects' right to explanation. By using a scalable, decentralised but coordinated approach we suggest laying bare general trends (that individual requests may fail to capture), highlight points where such rights and/or their enforcement need improvement and propose a concrete proof-of-concept for how algorithms can be explained in practice.

Co-creating an Algorithmic Interface for News Curation

As a first step in executing our proposed research agenda, we will apply the suggested approach to recommendation algorithms in the context of news curation. Nowadays, many people increasingly receive news from social media, where non-transparent algorithms decide which information they receive. This creates a 'filter bubble', which leads to less trust in traditional news sources and decreased media literacy, as it is harder to make critical decisions about what to read. As a result, misinformation (so-called 'Fake News') spreads more easily than ever. Making these algorithms more transparent and accountable can potentially increase media literacy as well as trust in traditional media. To achieve this, we adopt the following methodology:

1. *Desktop Research*: a cross-disciplinary mapping exercise of the relevant state of the art in both legal and HCI research. This will provide the basis for devising a detailed list of questions to be investigated in the empirical research phase (2), as well as input into the design research (3).
2. *Empirical Research*: a representative sample of popular online service providers will be identified, focused on providers building recommendation algorithms that are based on the profiling of users (e.g. social networks). The data will be gathered through surveys incorporating all the questions that emerged from the desktop research (1). The final results will be analysed by both legal and HCI experts so as to pinpoint the most problematic elements.
3. *Design Research*: Using a research by design approach, we suggest designing and evaluating a transparent algorithmic system. Through a sensitising activity, a small sample of users will be asked to document their experience of algorithmic systems in the form of a diary study. This will serve as input for two co-design workshops where we will define which elements of algorithms are to be made transparent. Based on the outcome of the workshops, we will prototype different versions of a user interface with varying ways of explaining the underlying algorithms. These prototypes will then be used in experiments to assess users' comprehension, acceptance and trust of the algorithmic system. In all phases, legal experts will contribute to the activities to ensure a close match with the results of the legal-empirical research.

It is expected that the learnings of the research will enable other interdisciplinary research teams to test various types of algorithms (e.g. in the context of Smart Cities, Healthcare or the Internet of Things).

Conclusion

As existing research on the right to explanation, algorithmic transparency and accountability has remained relatively theoretical, especially in law, with this position paper we aim to advance the scientific state-of-the-art in algorithmic accountability by filling two important gaps: (a) lack of substantial empirical evidence identifying key issues in the field; and (b) combine legal and HCI expertise so as to achieve adequate solutions (i.e. understandable and in accordance with the law). Notably, we propose an interdisciplinary research agenda with three important steps:

- Generate academically sound, empirical evidence on how the right to an explanation operates in practice;
- Apply an innovative and interdisciplinary methodology for empirical data-gathering of compliance with data protection rights;
- Create User Interface guidelines for algorithmic transparency and accountability, and a tangible prototype designed to explain algorithms to individuals;

References

1. Ananny Mike, Crowford Kate, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' (2016) New Media & Society, 1-17 http://dx.doi.org/10.1177/1461444816676645
2. Ausloos Jef, Dewitte Pierre, 'Shattering One-Way Mirrors – Data Subject Access Rights in Practice (2017) International Data Privacy Law, forthcoming.

3. Cramer Henriette, Evers Vanessa, Ramlal Satyan, Van Someren Maarten, Rutledge Lloyd, Stash Natalia, Aroyo Lora, Wielinga Bob, 'The effects of transparency on trust in and acceptance of a content-based art recommender' (2008) 18 User Modeling and User-Adapted Interaction 455-496 https://doi.org/10.1007/s11257-008-9051-3
4. Diakopoulos Nicholas, 'Accountability in Algorithmic Decision-Making' (2016) 59 Communication of the ACM 56-62 https://doi.org/10.1145/2844110
5. Eslami Motahhare, Rickman Aimee, Vaccaro Kristen, Aleyasen Amirhossein, Vuong Andy, Karahalios Karrie, Hamilton Keven, Sandvig Christian, '"I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds' (2015) Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems 153-162 https://doi.org/10.1145/2702123.2702556
6. Freek Evers, 'Handboek voor de Toekomst', *De Morgen.* Retrieved 30 November 2017 from https://www.demorgen.be/dossier/handboek-voor-de-toekomst/
7. Garfinkel Simson, Matthews Jeanna, Shapiro Stuart S., Smith Jonathan M., 'Toward algorithmic transparency and accountability' (2017) 60 Communications of the ACM 5.
8. Gurses Seda, Van Hoboken Joris, 'Privacy After the Agile Turn' in Polonetsky Jules, Tene Omer, Selinger Evan (eds.), *Cambridge Handbook of Consumer Privacy* (2017 CUP)
9. Kizilcec René F., 'How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface' (2016) Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems 2390-2395 https://doi.org/10.1145/2858036.2858402
10. Malgieri Gianclaudio, Comandé Giovanni, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation' (2017) International Data Privacy Law, forthcoming https://doi.org/10.1093/idpl/ipx019
11. O'Neil Cathy, 'The Ivory Tower Can't Keep Ignoring Tech', *The New York Times* Retrieved 30 November 2017 from https://www.nytimes.com/2017/11/14/opinion/academia-tech-algorithms.html
12. O'Neil Cathy, *Weapons of Math Destruction* (Crown Random House 2016).
13. Pasquale Frank, *Black Box Society* (Harvard University Press 2015).
14. Edwards Lilian, Veale Michael, 'Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?' (2018) IEEE Security & Privacy, forthcoming.
15. Tufekci Zeynep, York Jillian, Wagner Ben, Kaltheuner Frederike, 'The Ethics of Algorithms: From Radical Content to Self-Driving Cars', Centre for Internet and Human Rights. Retrieved 30 November 2017 from https://www.gccs2015.com/sites/default/files/documents/Ethics_Algorithms-final doc.pdf
16. Vedder Anton, Naudts Laurens, 'Accountability for the use of algorithms in a big data environment' (2017) 31 International Review of Law, Computers & Technology 206-224 https://doi.org/10.1080/13600869.2017.1298547
17. Wachter Sandra, Mittelstadt Brent and Floridi Luciano, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 76-99.