Algorithmic Decision Making in Online Sharing Platforms in extremis

Executive Summary

1. We are a group of researchers from diverse disciplinary backgrounds who have come together to try to understand how online spaces are being used by people in extreme circumstances. Our “A Shared Space and a Space for Sharing” project is one of several that form part of the EMoTICON network - an initiative jointly sponsored by the Partnership for Conflict, Crime and Security Research (led by the Economic and Social Research Council (ESRC)), Connected Communities (led by the Arts & Humanities Research Council (AHRC)), Digital Economy (led by the Engineering and Physical Sciences Research Council (EPSRC)) in partnership with Defence Science and Technology Laboratory (Dstl) and the Centre for the Protection of National Infrastructure (CPNI).

2. Our interest is in understanding the utility and applicability of online shared spaces for people who find themselves in extremis: for instance, people experiencing emotional distress or suicidal thoughts, caught in political revolutions, suffering from terminal illnesses, in need of a kidney transplant, or are exchanging information about “legal highs”.

3. We outline here three specific examples of such situations that highlight some of the functional, ethical and societal concerns that may arise from algorithmic decision-making. We also believe, however, that there may be benefits from algorithmic or other interventions in unmoderated social media forums, and we discuss these as well. Finally, we reflect on the on the benefits and perils of linking datasets for algorithmic decision making, and on handling errors caused by automated approaches.

Detailed examples and Reflections

4. Our first example is ‘Radar’, a Twitter based app launched in 2014 by Samaritans, a charity that provides emotional support to people experiencing distress. Radar was given the tagline, ‘turn your social net into a safety net’. Based on a simple algorithm using keywords [1], it allowed registered users to be alerted when a Twitter account they followed included messages that might suggest depressed or suicidal thoughts. Although the intention of the app was to support people experiencing emotional distress, and media and other responses to the app were initially positive, it proved increasingly controversial and was withdrawn nine days after its launch. In the period since, however, Facebook and other organisations have developed similar algorithms\(^1\) for the detection of suicidal users [2].

---

\(^1\) Although the algorithmic details of Radar’s and Facebook’s approaches are likely to be similar, Facebook’s initiative involves Facebook making contact with users that its algorithm identifies, whereas Radar made other Twitter users aware of distressed tweets.
5. The use of algorithms in such situations raises concerns about privacy, ownership of data, and confidentiality, as well as the possibility of errors (whether false positives or false negatives). The underlying premise of such approaches may also conflict with how people perceive the online space they are using. For example, an expression of distress might be a cry for help but could also be a means of managing despair; intervention in such situations, therefore, may be not only unnecessary but could be undesirable. If users view a particular online space as a place where they are safe to express emotion without interference, there is a danger that unwanted interventions based on algorithms may prompt them to retreat from online spaces, isolating them from the peer-to-peer support they may have built online. Other concerns about using algorithms in relation to emotions are to do with the apparent dissonance between the objective, mathematical and automated response based on the former and the distinctly human, subjective and nuanced phenomenon that is the latter. Indeed, an automated response triggered by the former in response to the latter may be regarded as inauthentic.

6. Our second example is our use of algorithms to detect bias in both news media and social media during the political conflict of the 2014 protests in Ukraine, also known as the Euromaidan Revolution [3]. Our technique builds on developing a model of word embeddings. This entails learning word occurrence probabilities to understand how language is laid out in short windows of words, around keywords or other words of interest. This allows us to zoom in on and identify words that are used in different senses in different texts. This software enabled us to show how pro-Russian news media might slant a news story differently from pro-Ukraine media. For example, our algorithm detected that Ukrainian sources used words such as “fake” or “non-legitimate” in connection with the Crimean status referendum. In the Russian Media, Ukrainian armed forces and the transitional government formed after the Euromaidan protests were strongly associated with words such as “fascism”. Algorithmic techniques such as these allow us to analyze how the same story is being seen from different points of view. One potential application of this might be to encourage each side to appreciate that during conflict situations there is another, potentially legitimate perspective on shared events. Equally, it can help to highlight propaganda, where words are subtly used in negative or positive contexts to manipulate public opinion in a particular direction. Algorithmic bias detection can be also used to ensure, or independently verify whether journalistic objectivity and balance are being maintained in traditional news media. It may help to distinguish between factual reporting of news and commentary on that news, a distinction that seems to be being increasingly blurred, especially when complaints are made about media bias, ‘fake news’, and stories are ‘tweeted’ in a few characters. However, we also show that the positions and slants of social media users can be predicted using the same techniques from their language use (e.g., by examining their tweets). This could have potentially serious consequences, for example, if repressive regimes start using bias detection techniques to identify and silence at scale users who are against official viewpoints. Users may themselves start retreating from public platforms, leading to the loss of a powerful avenue for self-expression.
7. Our third example looks at how user-identities from across different website domains can be integrated [4]. Our analysis shows that user-identities from well-established sites such as Facebook and Twitter can be much more trustworthy (in the sense of not sending spam or content that violate terms of service on a third website) than anonymous identities created on the site, or through email addresses not linked to third party sites. We developed a framework for inter-domain trust transfer, that moves us from the traditional weak and anonymous identities found on the Web to stronger identities that grow and accumulate trust over time, and across websites. A form of cross-domain identity linkage is becoming increasingly common, using for example the “Login with Facebook” (similarly Login with Twitter/Google+) buttons that are now being adopted by several websites. Such frameworks can be extremely useful in making the Web a safer place, and one that is easier and also more secure to use (since only one identity needs to be remembered rather than one per website). However, algorithmic trust transfer requires data to be shared across websites. This raises privacy concerns and other ethical issues. Identity-related information may be shared without users’ consent or active involvement. Such measures also make it harder for users to interact anonymously in circumstances where peer-to-peer advice is sought in relation to highly personal, sensitive matters where users may wish to adopt pseudonyms in one space and not another. Copying and sharing data across a wide range of domains also threatens the “right to be forgotten”.

8. Based on the above case studies, we recommend that entities using algorithmic decision making in extremis situations should:

1. Make explicit and easily understandable their assumptions about how users’ consent to be supported (on the basis of algorithms) is sought and given in online spaces;
2. Be aware that algorithms may be both potentially silencing and potentially amplifying deeply personal actions;
3. Recognise the sensitivity of the tipping point between support and surveillance;
4. Be aware that using algorithms, especially in relation to involvement in extremis situations – may be misconstrued as an inappropriate, insufficiently human response.

9. We conclude with two reflections on the use of data, and on managing the possibility of errors in algorithmic decisions.

10. On using data in unexpected ways: Algorithmically-driven decision-making often involves using datasets in unanticipated or new ways. Multiple datasets about individuals may be combined, or linked, in non-trivial ways to yield new insights – a potentially important justification and benefit of combining data. In many cases, however, personal user data are being used in ways that were not conceived of a priori. This may preclude explicit user consent. The benefits may outweigh the

---

2 For fuller discussion of these recommendations, see Brownlie, J. (under review) Emotional surveillance, digital outreach and the sharing of emotional distress online.
Written evidence submitted by “A Shared Space and a Space for Sharing” project (ALG0006)

absence of consent in some cases but combining data in this way may be problematic when the data relate to users in extremis. Should twitter data expressing suicidal ideation or methods for using legal highs be used in algorithmic decision-making? What are the implications of making erroneous judgement through errors in the data or in the algorithm with such vulnerable users? Questions such as these need to be addressed.

11. Handling the possibility of errors with algorithmic decisions: In many algorithms where machine learning is used to make algorithmic decisions, there will be some uncertainty involved – prediction algorithms are almost never 100% accurate or precise or complete in their answers. However, the extent of such errors, i.e., the level of uncertainty can be estimated in most cases, and this may allow the results of algorithmic decision-making to be interpreted appropriately, by considering the effect of Type-1 (“false positive”) and Type-2 (“false negative”) errors: would the decision change as a result of excluding the false positive, or including the false negative? If so, should the decision be changed, to “err on the side of caution”? An example of such reasoning is found in our recent work [5], which applied such methods to develop novel techniques for machine learning of nuanced labels about tweets made in the context of high profile deaths by suicide. Of course, in many situations, it is not clear what “caution” might be: For example, in the case of suicidal ideation is “caution” to be found in the preservation of privacy, freedom from undesired intervention or protection of life? Any action, and even inaction, in terms of not using algorithmic decisions, could have serious positive and negative side effects, and these should be carefully weighed on a case-by-case basis.

April 2017

References