Technology-mediated communication as co-production between humans and machines

Marinus Ferreira[1]

Abstract. As the use of technology-mediated communication becomes more sophisticated and comprehensive, there is a corresponding worry that its output is less that of a human agent using mute tools to deliver their own words, and more of a genuine collaboration between human and machine, with an accompanying loosening of our grip on exactly who or what is responsible for the content of the communication. Here I argue that we need not shy away from understanding such communication as co-produced by humans and machines, and how we can keep hold of ascriptions of responsibility for the content of the communication in the way we would for the communication from other collaborations. My proposal is to analyse technology-mediated communication as the product of an collective entity with sufficient internal structure to allow it to act in a coordinated and directed fashion. This combines and gives a novel application to recent work on algorithmic decision-making and collective responsibility, that being Kirsten Martin's work on the ethical import of algorithmic decision-making within organisations, and Stephanie Collins's work on group agency and group responsibility. We accordingly should not evaluate the products of such technology-mediated communication in terms of what either the humans or machines are capable of and responsive to, but what those people and resources united by this decision-making process is capable of and responsive to as a collective. I give special attention to how this agential group should be understood as handling the issue that the currently most popular tools for algorithmic decision-making treats its subject-matter as uninterpreted (in terms of arrays of numerical values, rather than the concrete objects that these values are drawn from and meant to represent), whereas communication necessarily must be interpreted for it to have referents and to be meaningful.

1 The contribution of algorithms

In the current state of play, the most widely used and studied technological means for authoring communication is the use of deep-learning algorithms trained on large selections of the appropriate kind of text. As such, I focus on this kind of technology, and will talk mostly about algorithms when I wish to discuss communicative technology. With that in mind, Kirsten Martin [4] has proposed that we understand an algorithm's contribution to decision-making as a matter of co-production between humans and machines ('co-production' is my gloss on her description of the interplay between human and algorithmic decision-making; Martin doesn't use the term). I will discuss her view on algorithms and accountability, and later in the piece apply this view to the case of technology-mediated communication in particular.

[1] Centre for Ethics in Public Life, University College Dublin, email: marinus.ferreira@ucd.ie

The issue that concerns Martin is a tendency to treat algorithmic decision-making as unanalysable: as being inscrutable to human oversight, or not available for inspection at all, and as not amenable to being corrected. This is for a variety of causes, most prominent of which is the sheer computational complexity of the algorithms. To cite an perhaps the most extreme example, the GPT-2 language model uses over 1.5 billion parameters [7]. But another cause, and one of genuine political concern, is that many of these algorithms as propriety and their owners bar others from inspecting them. The most discussed example is the use of the propriety COMPAS algorithm in sentencing in the judicial system, where neither the judiciary or those being sentenced are given the chance to see how COMPAS works [3]. There is also the issue of institutional inertia: the fact that the use and development of algorithms in decision-making is a going concern, with resources that have been committed to it and processes that are underway that depend on it, means that halting in order to inspect and perhaps overrule the outputs of algorithmic decision-making would introduce unpalatable delays to the whole system which makes use of these outputs, and as such the more use is made of algorithmic decision-making within some domain the less appetite there is likely to be for second-guessing it. And finally, there is the simple reticence by institutions to be placed under oversight, not at all restricted to those institutions making use of algorithmic decision-making. The combined effect of these factors is to downplay the very real possibility that the algorithm can introduce mistakes in the reasoning in question. As a result here is a resulting and pernicious tendency of magnifying the responsibility born by humans (both those who are part of the decision-making process and who suffer the consequences of the decisions) and of diminishing the responsibility of the algorithms.

Martin makes an extensive criticism of all of these factors that work against the oversight of the role of algorithmic decision-making. Her focus is on introducing tractable ways of responding to the errors that algorithms can introduce to a piece of decision-making. Just like humans, algorithms bring different capacities to the table, and are likewise liable to make their own mistakes or biased decisions. The tendency that Martin is working against is for accountability for the errors in algorithmic decision-making to wither away in the face of the real or imagined openness of the algorithm to oversight; by tackling the prospect of errors made by algorithms head on she attempts to correspondingly make the prospect of having accountability for the decisions in question a real prospect as well.

Martin's engagement with possible algorithmic errors takes place across a number of fronts. One important aspect is that the interest we have in using an algorithmic decision-making process depends on what she calls the 'social embeddedness' of the process: the context for it, including how it could have been done differently and what its

different results would have been. This, together with reflection on the actual outputs of algorithmic decision-making, means that there is a trade off of the burden of responsibility between the developers of the algorithm and its users. Where the errors that an algorithm may introduce are something the user can see and correct, the responsibility for the outputs shifts from the developers to the users; where this is not a real prospect, the algorithm is treated like a black box and responsibility lies squarely with the developer; where the errors can be seen but not corrected by users, or could be corrected by users but are hard to find, then the algorithm is not considered useful. Martin articulates this in terms of algorithms being evaluated on two dimensions, social embeddedness and reflection, and that the use of an algorithm is least likely to be pernicious when that use scores highly on both of these dimensions, that is, when errors can be seen and corrected by users, giving them agency over the use of the algorithm.

As we can see, dealing with the prospect of algorithms introducing errors does not mean that we obviate them of their interest, provided appropriate allowance is made for such errors. Just as in the human case, we can find the contribution of algorithms to a decision-making process worthwhile even when acknowledging the prospect of introducing errors, because the capacity of a group of fallible agents often outstrips that of an individual fallible agent. I will give a small example of this later in the piece. As such, Martin offers a framework for judging what decisions can appropriately be left to algorithms and which need human intervention to be put into effect. Her proposal is to highlight the two dimensions I discussed above, the social embeddedness of an algorithm, and the reflection involved in it. Neither of these are features of the algorithm; they are features of the algorithms use. In the rest of the paper I introduce another framework, that of group agency, in order to explain how human-machine coproductions, can serve to give both the social embeddedness and the reflection that Martin compels us to aim for.

2 Human-machine collaboration as an agential group

Stephanie Collins in her recent monograph [1] provides an analysis of different kinds of groups and the duties they are liable to have. I use her notion of agential groups, aggregations of individuals and resources joined with a decision-making process that allows the group to be jointly responsive to and mobilised towards some chosen end, and apply it to the collaborative authorship of technology-mediated communication. This is meant to put a sharper edge to the framework on judging who is responsible for what offered by Martin, by harnessing recent advances in our understanding of group duties and collective responsibility to the case of human-machine coproduction. Of particular interest here is that we have a clear avenue to understanding the internal structure which is meant to give the human-machine collaboration the standing as an agential group, because this structure is a matter of explicit design in the implementation of technology-mediated communication.

Collins offers a framework for describing when a group has sufficient internal structure to be treated as a responsible agent. My suggestion is that if some individual or group of people uses technology-mediated communication, the process through which they do so (presumably, the process through which they give the appropriate input to some algorithm in order to receive the appropriate output) would count as a decision-making procedure in her sense. She concentrates on moral responsibility, because establishing when and how groups can bear moral duties is the most difficult and most pressing task in the literature she is engaging with, but that focus need not concern us here. It is enough to note that a group that can bear moral duties can bear responsibility for the products of technology-mediated communication as well. Her criteria for a group being able to bear group duties are as follows:

(i) there is a procedure that is operationally distinct from the analogous procedure of any other agent, which is robustly able to generate rational (that is, not mutually contradictory) decisions, based at least partly on moral considerations, on the following:
  (a) which goals a group will pursue, and
  (b) roles that are jointly sufficient (absent defeaters) to pursue the goals in (a), and

(ii) the roles in (b) are all assigned to agents; and
(iii) each agent in (ii) has systematic influence, licensed by the procedure in (i), over the decisions in (i);
(iv) each agent in (ii) has committed (defeasibly) to abide by the procedure's decisions and this is a matter of common knowledge amongst them; and
(v) each agent in (ii) is able to receive sufficient information about the procedure's decisions for them to abide by it.

It should be clear that these criteria are met by a group who is able to produce a piece of communication: a team tasked with writing a piece of copy, a committee drafting a piece of policy, collaborators jointly producing a scientific paper, etc. Let us discuss the criteria (i)-(v) in broad outlines. A group collectively writing a piece of communication has some procedure fitting with (i), which is the procedure they are using to settle the content of the communication, even if that process is nothing more than purely ad hoc face-to-face discussion, that is enough to count. Similarly, the roles in question for (ii) need not be in any way strict, nor do they even need to be articulated, since (ii) is a condition to say that there is not just some potential group (not yet assembled) who is engaging in the task, but an actual group that is already engaging in the task, and (iii) is there to establish that the actual people involved in the group are not epiphenomenal to the process but have an influence in the form and content of what the group is doing. By the same token, (iv) and (v) are there to say that the process in question is sufficient for and successful at guiding the individual members in engaging in the group's task. So, for example, when a task force is writing a statement concerning a new product their company is offering, that counts as a group agent by Collins's criteria in the commonplace scenario where that statement actually does get written, each member of the task force gives some input that is not just ignored, and nobody outside of that task force writes any of the copy. In the case where there seem to be two bodies, say, the leadership of a company and the task force who writes statements, they are for Collins's purposes just one group, because it is these two groups working together (the leadership providing a brief to be met, the task force producing the actual copy) that counts as the group that acts together and is collectively responsible for the action.

Taking on board what was said above about the ways that algorithms and other technology can be seen as contributing to processes, my contention is that these kinds of groups supplemented with the tools required for technology-mediated communication also meet Collins's criteria for being a group agent. This may seem to assume that algorithms are examples of genuine intelligence which are appropriately treated as on a par with humans, but instead what is happening is that the question of what the extent of the artificial intelligence is gets sidestepped. The Collins criteria makes no reference to the capacities involved in participating in an agential group,

and this is right and proper since it is commonplace that a group does not mobilise all of the capacities that an individual agent may have. It is not that to play your part in a group process all of your aspects of human agency needs to be devoted to that group; instead, you only need the capacities that are called upon by the group's processes. By the same token, while under the Collins criteria talks about the decision-making procedure at use in a group, it does not require each member of the group to be able to design or direct such a group-level procedure, only for their activities in the group to be responsive to it. This means that provided that the use of an algorithm fits within the group's decision-procedure, we need only evaluate the output of the process—the rest is a distraction.

We can arrive at the same conclusion by sidestepping what the extent of genuine artificial intelligence is through using the fact that groups contain not just people and processes but also material resources, as Collins allows (following Epstein [2]). This means that when we evaluate the capacities of the group, we need not divide the group into agents and non-agents and see what the capacities of the agents taken together are, but (again) need only evaluate the capacities of the group taken as a while, situated as it is where it is with the resources that it has. This approach may suit someone more hesitant to consider the contributions of algorithms as analogous to human contributions than Martin or I am; for present purposes nothing hinges on this fact, though of course there is great and widespread interest in the question of how to handle the contributions of technology to human processes as more than mute tools or pieces of equipment, and this paper is a contribution to that larger question as well.

That the use of algorithms for technology-mediated communications would count as the operation of a group-level decision-procedure, just in the way that face-to-face collaboration among humans would, can be seen from the fact that, by way of being a computer program (or suite of programs), the use of the algorithm requires following a specific process. If the group is set up to make use of that algorithm, then the decision-procedure in question includes these programmatic procedures (i.e. to write a statement, the group will give a writing prompt to the algorithmic program, which is used as the first attempt at the task). This is more explicit and more structured than the way that a human collaborator may contribute to the authoring of a text, and obviously being more explicit and more structured makes it correspondingly more suited to featuring inside of a decision-procedure as required by the Collins criteria.

In turn, the use of the algorithm in such a context would help move it along the two dimensions Martin highlights, social embeddedness and reflection. The social embeddedness comes automatically with the use of the algorithm within a somewhat defined group structure such as the Collins criteria require. This is because the algorithm is then placed in a concrete context of use, with specific aims in mind and other parties engaged with it in a way that makes the actual and expected products of the algorithm more tractable. The increase in reflection would need to accounted for case-by-case, but use within such a group-level decision-procedure would aid it. This is because the algorithm is not just taken to run isolated from other parties, but because it is part and parcel of a wider context which in turn guides expectations about what it should produce, in turn guiding evaluations of what it does produce. As such, we move from the use of an algorithm as a black box process to what Martin identifies as a far better use, where its users can assess the outputs and take steps to mould them in the ways they require.

3 Interpreting technology-mediated communication

There is a deep worry with machine communication, in that the trend of recent artificial intelligence technologies, especially deep learning algorithms, is to increasingly treat its subject-matter as uninterpreted: medical diagnosis algorithms deal with sets of numbers, not features of patients and diseases, and natural language processes increasingly don't treat communication as linguistic items but as statistical relationships across entries in a database. This threatens our notion of authorship (both of decisions and of communication) because we cannot make sense of what is involved with understanding a piece of language without interpreting of the outputs: not just seeing them as numbers or other formal symbols, but having a mapping from these numbers onto tangible situations, and judging the outputs in terms of what they say about these situations. The less contact there is between the technological methods and the interpreted domain they are applied to, the less of a grip we have on the content of its products. I take this to be the deepest problem with technology-mediated communication (and a very important problem for algorithmic decision-making in general). This is especially important for the case of authoring communication, because as competent as machine processing of language may be at various tasks, there is no prospect of current technologies treating pieces of language as referring to tangible situations (and the current raft of advances in natural language processing has come about after largely abandoning this as an aim). As such, this problem regarding interpretation is the most prominent strand of critique of statistical natural language processing methods among linguists [5], which we can easily extend to the case of technology-mediated communication.

Two different responses to the interpretation worry present themselves. Firstly, we can try to a distribution of labour, by carving out a role for machines in the collaboration which does not require any interpretation of the communication it deals with, and leaving humans to deal with the relations between the linguistic items in the communication and their referents. In the words I used at the beginning of the piece, this would restrict machines to the role of mute tools, insofar as this is possible when producing communication. Secondly, we can try and explain how machines can be involved in the production of pieces of language it has no prospect of understanding, and lean into the suggestions I've made about humans and machines co-producing communication. I defend the second approach.

As has been argued above, we can take the human-machine collaborations as an agential group which is the target of our judgements about the suitability or not of some piece of co-produced communication. The point of technology-mediated communication is that in many instances the capabilities of the human-machine collaboration outstrips that of either human and machine, and we leave that by the wayside when we silo responsibility for different aspects of the tasks between humans and machines. By using the standing of a human-machine collaboration as an agential group, we do not need to judge a piece of communication twice, first as appropriate for algorithmic capacities and then as appropriate as a piece of communication. We need judge it only once on its merits as a piece of communication produced by the collaboration.

Let us return to the interpretation worry. An algorithmic process may struggle with giving tangible interpretations to pieces of text, but any human who is a competent speaker of the language in question does not. Taking again the example of GPT-2, there is a simple and much-publicised example of technology-mediated communication produced with the help of GPT-2, a mock news report generated

from a human-given writing prompt of the discovery of a population of English-speaking unicorns in the Andes [6]. While GPT-2 displayed an uncanny grasp of the conventions of that kind of writing, and managed to passably imitate the performance of a competent language user in a number of respects that algorithmic processes often struggle with (for instance, in being able to make passable use of the same piece of background information at relatively distant parts of the text), a speaker of English would be surprised by a number of infelicities in the text, such as the description of unicorns as having four horns (or of unicorns being produced by a meeting between humans and unicorns, or that meetings of this kind are said to be quite common in South America). This is of course because the interpretation of the word 'unicorn' is of an animal with one horn, not four. It turns out that the enormous arrays of numbers assigning probabilities to the collocations of text strings that underlie the performance of GPT-2 did not manage to reproduce text respecting this common understanding. Undoubtedly, if a human were to take the text up to that point, amend it to referring only to a single horn, and feeding it back to GPT-2 (in effect, making the text up to the amended point a larger writing prompt for repeating the task but with the human correction), then the algorithm would produce a text with the same virtues as it had before, but without this particular error. Maybe it would produce further errors, maybe it wouldn't, but the point is that the human-machine collaboration has a capacity (in this case, the prosaic ability to not contradict our usual understanding of infrequently used kind-terms) that the algorithm on its own does not.

The above-described process of human amending of machine output in order to smooth it out to meet the expectations of a human audience is only a toy example, but it suffices for demonstrating the point I want to make here. Given what was established earlier about human-machine agential groups, there is nothing in principle which stops us treating this back-and-forth between human and machine members of a co-productive partnership the way we would a similar interplay between different human members of a collaboration. We also need not commit to the particular kind of contribution the different members of the co-producing collaboration make to the final product, because what matters is that there is some effective decision-making procedure that unites them. Again, by treating the co-producers as an agential group we side-step having to make evaluations of the capacities of any of the individual members one-by-one, we need only consider the capacities of the group taken as a whole.

4 Conclusion

The difficult part of the work being done here is not to describe the technical process involved in managing human-machine co-production, but to give a tractable and informative framework for how to conceive of the issues regarding responsibility and control in the cases where algorithms and the like contribute to technology-mediated communication. To do so I introduced two different philosophical frameworks, Kirsten Martin on designing ethical algorithms and Stephanie Collins on group agency, and harnessed them both in novel ways to the question of human-machine co-production of communication. I discussed how Martin highlights that to give account of the contribution of algorithms to human processes we need to be clear about how errors that the algorithms may introduce feature in the running of the process, and the two dimensions of the social embeddedness of and reflection upon the working of an algorithm are good avenues for doing so. I describe how these two dimensions are developed through the incorporation of an algorithm or similar technical process into a group-level decision-procedure for the authoring of communications, and described how this allows us to evaluate human-machine co-production of technology-mediated communications.

REFERENCES

[1] Stephanie Collins, *Group Duties: Their Existence and Their Implications for Individuals*, Oxford University Press, Oxford, 2019.

[2] Brian Epstein, *The ant trap: Rebuilding the foundations of the social sciences*, Oxford University Press, USA, 2015.

[3] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin, 'How we analyzed the compas recidivism algorithm', *ProPublica*, (23 May 2016).

[4] Kirsten Martin, 'Designing ethical algorithms', *MIS Quarterly Executive*, **18**(2), (2019).

[5] Joe Pater, 'Generative linguistics and neural networks at 60: Foundation, friction, and fusion', *Language*, (2019).

[6] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever, 'Better language models and their implications', *OpenAI Blog*, (14 February 2019).

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, 'Language models are unsupervised multitask learners', *OpenAI Blog*, (5 March 2020).