# AI in Culture & Society

AI Culture Lab - Waag
Tom Demeyer
v.1

waag
technology & society

## introduction

Of course, where else? In this piece we'll be looking at AI technologies mostly from the perspective of the humanities. More precisely, we'll be looking at what the use of AI can teach us about ourselves and our communities, and how society and public policy might benefit from the advent of more and better instruments to set priorities, understand preconceptions, suggest courses of action and evaluate results.

The highly visible, and much discussed applications of AI in advertisement and surveillance have a huge impact, of course, on the structure and functioning of our societies, and not necessarily for the better[1]. In fact they help spark the sentiment of a 'race' we have to participate in, or have already lost, without much (or any) consideration for the goal or direction of that race, other than achieving technical supremacy (and world domination). With the advice given to policy makers: "join the race or get left behind," we're missing the point altogether. A race towards more economic growth and unsustainable consumption or more effective state control is not a race we as citizens need to participate in. Instead we'll be looking at these technologies with a designing purpose; where and how we can exploit the sometimes almost unreasonable effectiveness of some of these techniques to work on better versions of ourselves, our institutions and our societies. Value-driven, open and inclusive design will help to define 'better' to have an operational meaning in the relevant application context and for the relevant stakeholders.

There's no clear picture or map of where current AI technology is deployed, in which domains and with what results or effectiveness. We cannot draw a clear line between traditional business intelligence, statistical analysis, data-driven decision making and what we currently mean by AI. It is clear, however, that huge strides have been made in various vertical domains in the last decade. In military applications, we presume, in geological surveys, financial services, in medical diagnosis, and, not to forget, in assistive applications (from translation and Alexa to sports- and weather 'journalism'[2]). The absolute supremacy of deep learning technology in gaming contexts would also suggest a huge potential in applications for international diplomacy and conflict management, although we don't see any evidence as of yet…

[1] https://crackedlabs.org/en/corporate-surveillance

[2] https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html

One potential that we do not yet take advantage of is the power of AI to study our collective selves. Not in the sense that AI is comparable to human intelligence, it is not, but in the sense that we can 'freeze' a small section of our culture and lift it out into a model, enabling us to prod, probe and study that little bit of our existence in ways that we have never before been able to, potentially gaining insights that would otherwise be much harder to achieve. This we will explore in the rest of this paper.

## culture as bias

Whenever applications of AI are discussed that affect people directly, such as predictive policing, job applications management or repeat offence prediction (COMPAS[3]) there's talk of bias. Usually the systems turn out not to be neutral with regard to ethnicity or gender. Amazon's recruiting software perpetuated male dominance in tech jobs, penalising women for not having been hired in the past, effectively[4]. Defining 'neutral' is a challenge in itself, but we can conceive of a statistical definition regarding a desired behaviour in a particular context. The point of course is that, as the training data will not be neutral, we would need to 'skew' the algorithm, either in the learning phase or in the inference phase. Even if we could isolate the parameters under consideration, no technique is foolproof, and there is usually a trade-off between accuracy and fairness[5]. Indeed, even when the unfairness is obvious and easily detected it turns out still to be very hard to ameliorate because it is impossible to understand interdependances between the multitudes of attributes captured in a typical model.

When corrected for gender bias, the Amazon system did not perform better than random and was abandoned. Of course the problem is far more insidious when systems are used where unwanted bias is not obvious and not (easily) detected.

There are many articles and publications on bias and how to avoid it. Bias is categorised, and recipes supplied to address each class. The problem with this 'how to prevent machine bias' approach is that, in general, the desirable classification behaviour that we

[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[5] https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

train our models for is indistinguishable from undesirable bias, in a technical sense. Classification or grading is desirable on one axis, detrimental on the other. How do we know which axes there are, which are relevant, which desirable and which not? Typically, the inner workings of a trained model are too subtle to understand directly; we cannot intuit the mechanism of operation and correct the model on that level. If, during training, we try to anticipate all possible dimensions and identify and prevent the undesirable ones from expressing, we'll end up building a rule-based system which we now know cannot hold a candle to deep learning networks in many applications.

This is not to say that we cannot or should not bother ourselves with avoiding unwanted bias; obviously we should do our utmost to build models that are effective at the job and do so in a way that is consistent with our moral code.  In all systems that affect people directly, but especially where they touch on (constitutional) law or basic human rights, we can never assume, however, that the system is neutral in any way but always reflects both the culture in which is was designed as well as the circumstance in which it is used.

## context, context

When we think of 'culture' we could, for the sake of argument, think of this as a huge set of dynamic interacting and interdependent preferences, biasses in the above sense, for all intents and purposes. We could be talking about Western Culture, or maybe something less huge as in the "KLM culture" vs. "Air France culture".  Or even smaller as in "Amazon's hiring culture over the last decade".

The point being that we can, if we allow ourselves, see the training of a model as capturing a chunk of culture, freezing it, and storing it in a queryable framework for later use. The context from which we took the 'chunk' remains behind, and the chunk becomes static (in the sense of 'frozen', not 'noise'). When we later 'perform' the model, present it with relevant inputs, it expresses in a new context, maybe very similar, maybe not so. There are many ways in which context can differ;  interpretatively, as in culture, technically, or in purpose,  application, for instance.

Models are, in this sense, not universal; just imagine an U.S. trained COMPAS being used to judge repeat offence chance of people in a Norwegian court. Or, closer to reality, a U.S. trained model to detect offensive content in Dutch online expression (where US prudishness regularly exasperates artists & publishers). This mismatch may also occur across time, of course, where a changing morality causes a system to be no longer effective where it previously performed acceptably.

We would not refer to this kind of issue as a problem of bias, although in non-obvious cases it is not very likely that there is a fundamental difference in effect between a model / context mismatch and bias issues. We could say that when the model is performed, the application context as well as the choice of inputs -the question that we ask- inflict a further set of both desired and detrimental biasses that we need to be aware of.

Of course, in the obvious cases we'll catch the mismatch, or won't even consider the application. But, as with bias, the systems could produce non-obvious detrimental results from subtle mismatches in context that we are not aware of.

## culture probes

If we, in a tour de force of suspension of disbelief, allow ourselves to reduce culture to a set of interacting preferences, and to define a preference (a bias) as a probability distribution on a particular axis (where a normal distribution would be the absence of bias on a particular parameter), we can see a many-dimensional matrix taking shape. This matrix would provide us with a means of probing, measuring and studying the encoded chunk of cultural practice using mathematical and computational techniques. Of course, identifying the axes and distributions in the first place is going to be more than a challenge. Starting with a very small domain may offer some insight and suggest some approaches to study both the model and the 'culture' it encodes. Apart from a strictly technical analysis, through successively and selectively exercising the model with (artificial) input vectors we might try to find the emergent dimensionality of the encoded domain, or even try to identify the most prominent axes affecting the results, as well as their distributions. Since the original context has been 'cut off' and the model is a frozen representation of the domain we're investigating, and it does no longer adapt to use as the system it represents originally did, there will be quite some caveats in drawing any conclusions about that original domain from these exercises. Nevertheless, as with studying liver cells in vitro, we can still gain a lot of insight into the systems.

It also becomes interesting to speculate about ‘reverse AI’, where we have a real input, or set of inputs; we have designed, or there is agreement on an ideal result, given these inputs, and we try to find or calculate the ‘training data’ the would produce the model that generates these outcomes. We then ask questions like ‘what culture would produce a good or optimal x or y, given these inputs’. This would then allow us to evaluate and possibly adjust current practices underlying the domain.

This culture probing is highly speculative, of course, but the experiments may lead to increasing our knowledge of the way in which the models encode the domain they are trained in and may also give us a better way to differentiate between wanted and unwanted bias, even if it will be very hard to ameliorate the one without compromising the other.

The advent of the use of AI has drawn attention to way we work and made explicit, in some cases, the underlying assumptions in our behaviour; not only the biasses of the models we have trained. In the humanities, we already profit from the advent of AI in the sense that we have now have a chance and an excuse for a renewed focus on the way we make decisions and on the hidden mechanisms in much of our cultural practice. We should ride this momentum and thereby better understand and design both the technology and the society  that we live in..

waag
technology & society