"Put it in the bin" Mapping AI as a framework of refusal

Garfield Benjamin*
Solent University
Southampton, UK
garfield.benjamin@solent.ac.uk

Abstract

This paper presents a sketch for mapping AI along relational lines, centring those who are most affected by an AI system, identifying asymmetric power relations, situating systems in wider discourses and narratives, and thereby locating points of resistance, refusal and the redistribution of power.

1 What does it take to send an AI system to the trash?

Many AI systems should not exist. They require resistance. For example, police use of live facial recognition in public has received much needed resistance across the world, using a range of approaches. In the UK, a lawsuit was brought against South Wales Police, resulting in its use of facial recognition being declared illegal (but only after appeal), while the Scottish Parliament received evidence from academics (myself included) [16], think tanks and advocacy organisations which persuaded the Justice Sub-Committee to pressure Police Scotland into removing facial recognition from their future plans. Meanwhile, as its use nationally continues, think tanks such as WebRoots Democracy [17] have engaged more directly with marginalised communities in the UK to demonstrate the racist discourses and impacts of the technologies. Various of these sources of refusal, along with other resources from researchers, think tanks, advocacy organisations and the press, were used to construct the example mapping of police use of facial recognition in this paper. There are calls from many different angles for specific AI systems to be defunded or disbanded, and in some specific areas (often at devolved or local levels) they are gaining political weight. In a Twitter conversation on the London Metropolitan Police's claimed exceptionalism for continued deployment, Hannah Couchman of Liberty human rights organisation said simply "put it in the bin" [8]. But how do we make a case to send AI to the trash? What tools can we develop to map the different approaches, issues and narratives that can persuade decision-makers to abandon this or any other inequitable or unjust AI?

There is a need for conceptual tools appropriate to engage with different audiences: the researchers and companies who develop an AI system; the funders enabling its development and defining its priorities; the governments (and other organisations) who buy and use it; the public who accept or resist it, particularly the most affected or marginalised communities who may need tools with which to speak truth to power, and by extension the media who represent it to the public, together forming the contextual norms surrounding it; and the policy-makers who regulate it, converting norms into enforceable law to protect rights and enact justice. In this paper, we propose one such conceptual tool, a mapping of AI as a sociotechnical assemblage that focuses on affected communities and can offer a framework for identifying criteria for refusal and resistance to inequitable AI systems. An example mapping is presented focusing on policy use of facial recognition, and further questions are presented for future work.

*https://digitalcultu.re @g8enjamin

ResistanceAI Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

2 Mapping AI

A preliminary sketch for how to map the sociotechnical assemblages of AI was applied to drones [1], and the process can also be informed by, for example, Kate Crawford and Vladan Joler's Anatomy of an AI [9] (which focused on labour and ecology). To develop the toolkit and process further, we need to outline the principles, possible criteria for refusal, and structure of mapping the power structures and impacts of AI.

2.1 Principles

Mapping AI as a tool for resistance should follow several key principles.

Relationality is fundamental to understanding asymmetric power structures. This can include a combination of justice (for resistance) and care (for building new equitable structures) ethics but also leans on performativity [2] to understand the social construction of norms through different actors in asymmetric contexts. Major AI ethics guidelines largely ignore social embeddedness and care [11] but, along with justice and relationality, these are well established in intersectional STS [3,15].

Centring the margins is perhaps the most important step in building a tool for resistance, building on Black Feminism [13], its application to AI [4], data [10], and starting the discussion of broader tech from marginalised intersectional perspectives [12].

The right to refusal is a principle and aim that guides this project. The ability to say no needs to be raised at all levels, from refusing the use of data [7], to labour, to ecology to culture to harms in practice and the perpetuation of oppressive narratives. Refusal can be individual and/or collective.

Across these principles, we elevate intersectional approaches including (but not limited to) critical race, feminist, queer, trans and disability theories.

2.2 Criteria

What measures, what factors, can be presented to make a case for binning AI? Current legal requirements around, for example, the data protection and privacy rights issues that are entwined with AI often focus on proportionality. But this is a deeply flawed measure that can always be used to justify what those already holding power wish it to. Qualifying rights, harms and justice in this way only works by closing off sections of the context, by excluding those who are most affected. But what other measures can mapping AI provide to help us resist?

Validity of research needs to be questioned. There is a need for a shift towards greater burden of proof, rather than disproof as has been the case in, for example, facial recognition [5]. This involves transparency of the research process rejecting publication of selective best-case outputs, as proposed for clinical trials giving the same scrutiny for AI as for drugs [14]. This can also be aided by contextualisation, making visible the framing of research and resisting the false objectivity of data and models that underpins much technical research. A relational map can help expose these pitfalls of positivism, and highlight the persistence of discredited and racist practices.

Cost can be a major tool. Vast swathes of public and private funding are being funnelled irresponsibly into harmful AI research. For example, there is no compelling case that police use of facial recognition is a justifiable use of public funds. Mapping where money goes (usually lining the pockets of already wealthy tech companies) can show where it can be better spent, for example to much-needed public services. Mapping should highlight not just power but financial flows as a tool for exposing corrupt practices and influence, essential to making visible how the entire ecosystem functions [6]. It can also demonstrate flawed assumptions in cost effectiveness by moving from cost-benefit to cost-harms analysis. This can provide a clear case for where AI needs to be defunded.

Narratives of AI can be harmful or misleading. There are many blurred definitions and mis/representations of what AI is or can do, but by shifting the focus onto the broader social interactions and power structures - rather than technical objects or marketing rhetorics that often displace agency and in doing so divert responsibility [1] - a contextual and relational narrative can be used to target resistance at the appropriate levels of decision-making. This may be materials, design, data, operation, regulation, broader political or business interests, or issues with the underpinning fields of research.

2.3 Structure

Mapping AI as a sociotechnical assemblage acts as a social audit of AI. It allows us to construct different forms of objections, and the process itself can be a means of elevating different voices and narratives, with space for community or wider public consultation. Roles, components, entities and narratives can be placed on the map using information gathered indirectly from research, advocacy groups, the press, or directly in conversation with specific (affected, labour, research or user) communities. The proposed method of creating a map is a series of concentric layers, as follows:

Centre: those affected by an AI system, particularly those with less power/choice. This is those whose data is used, those upon whom the decisions of AI are enacted, and those who are otherwise affected by the use of the AI system;

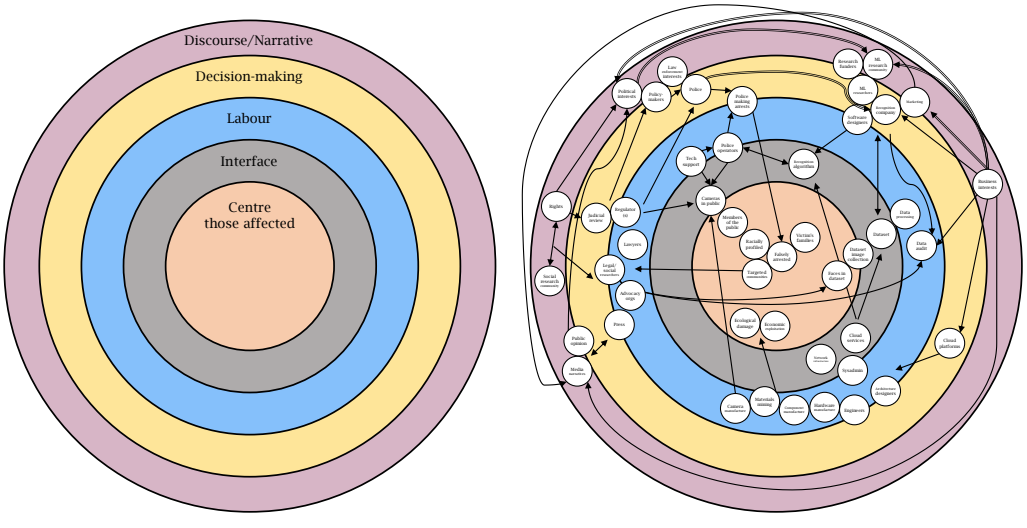
Interface: the AI as a technical object, including its constituent technologies and their interactions. The boundaries of this layer are where the interactions occur, and the placement of this layer between those affected and those using/designing/deciding the system is a purposeful and political gesture to highlight the anonymising effects of AI as an interface for social issues;

Labour: the operators or users of the interface layer, the designers, coders, engineers who build it, the miners of materials and manufacturers of physical components, and the sysadmins and others who maintain the system - in short, all those whose labour contributes to the designing, building and running of the AI system. Placement relative to harms and decision-making can show relative exploitation or reward;

Decision-making: moving outwards, this includes the relative power of decision-making by researchers, designers, businesses, organisations, clients, funders or policy-makers, that defines the development and deployment of AI;

Discourse/Narrative: also known as the principles and influence layer, this is the systemic forces that define the context in which the decisions are made. This may be fairly consistent across different AI - particularly in more theoretical sub-fields - and includes the press, business interests, political (including military) interests, funding body priorities and the shaping of the research community.

Each individual, group, role or interest is located on the map and linked to other elements to show influence via directional arrows. Additional information can be provided as overlays to show specific types of harm, include links to narrative components or discussion, or show different types of influence (such as funding streams). Figure 1 shows this method of mapping with: (a) a generic template; and (b) a specific example - police use of live facial recognition in public (in the UK).



(a) General template

(b) Police use of facial recognition

Figure 1: Mapping AI.

3 Discussion

Uses The aim of this tool is that it acts as one way to identify power asymmetries in AI systems, in order to empower acts of refusal. This follows the data feminist approach of identifying and then challenging power [10]. The mapping process presented here can be used to identify harms and power asymmetries, and in doing so highlight systemic problems. Points can then be found where interventions can be made - via collective action from research, industry and civil society - in order to counter harms and redistribute power along more equitable lines.

Open questions The method presented is an evolving process, and is only one possible way of identifying and visualising the inequitable power flows and impacts of AI systems. For example, further attention could be given to mapping the impacts and issues with theoretical research. Few mappings will be able to be fully exhaustive. Further development could include ways of overlaying narratives of the harms, weighting influences, and filtering forms of influence, all of which could be added as options in an interactive tool. Additional measures can also be developed and visualised that provide concrete and narrative societal measures against which unequal 'benefits' could be countered.

I started this paper by asking "what does it take to send an AI system to the trash?" But by system we often mean an entire sociotechnical assemblage that requires systemic change. The ways AI escalates existing inequities and injustices can highlight this need for broader societal change. Facial recognition shows the long history of racist policing; financial algorithms show economic inequities and redlining; search result biases show persistent problems with representation. By contextually situating AI in these existing narratives we can formulate a wider case for collective action.

This paper hopes to offer a way of visualising the asymmetric power structures and harms of AI systems. It takes a relational and contextual approach that centres those who are most affected, acknowledges unequal decision-making structures, and embeds these decisions within broader interests and discourses. The map can be used to identify in a structured way the relationships that need addressing to redistribute power and counter harms, with the aim of acting as a conceptual tool for empowering and enacting refusal of AI by affected communities. The tool can also be applied to other technologies beyond AI, and it can also be used to highlight broader systemic problems in social structures that enable the continued development of unjust and inequitable AI. This is not a fixed tool but a constant and contextual work in progress.

References

[1] Benjamin, G. (2020) Drone culture: perspectives on autonomy and anonymity. *AI & Society*, 1-11.

[2] Benjamin, G. (2020) From protecting to performing privacy. *Journal of Sociotechnical Critique* 1(1), 1-30.

[3] Benjamin, R. (2019) *Race After Technology*. Polity.

[4] Birhane, A. & Cummins, F. (2019) Algorithmic Injustices: Towards a Relational Ethics. *Black in AI, Neurips2019*, 1-4.

[5] Buolamwini, J. & Gebru, T. (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *FAT19*, 77-91.

[6] Carmi, E. (2020) *Media Distortions: Understanding the Power Behind Spam, Noise and Other Deviant Media*. Peter Lang, p.159.

[7] Cifor, M., Garcia, P., Cowan, T.L., Rault, J., Sutherland, T., Chan, A., Rode, J., Hoffmann, A.L., Salehi, N., Nakamura, L. (2019). Feminist Data Manifest-No.

[8] Couchman, H. (2020) Put it in the bin. More detailed thoughts in due course. *Twitter* 10:06 AM Aug 19, 2020.

[9] Crawford, K. & Joler, V. (2018) Anatomy of an AI system.

[10] D'Ignazio, C. & Klein, L. (2020) *Data feminism*. MIT Press.

[11] Hagendorff, T. (2020). The ethics of Ai ethics: An evaluation of guidelines. *Minds and Machines*, 1-22.

[12] Hoffmann, A.L. () Data, technology, and gender: Thinking about (and from) trans lives. Pitt, J. & Shew, A. (eds) *Spaces for the Future*. Routledge, 3-13.

[13] hooks, b. (1984) *Feminist theory: from margin to center*. South End Press.

[14] Liu, X., Cruz Rivera, S., Moher, D., Calvert, M., Denniston, A. & The SPIRIT-AI and CONSORT-AI Working Group (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine* 26, 1364–1374.

[15] Noble, S. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[16] Scottish Parliament (2020) [Report] Facial Recognition: how policing in Scotland makes use of this technology.

[17] Chowdhury, A. (2020) Unmasking Facial Recognition. *WebRoots Democracy*