1. Introduction

Automatic Authorities

Automated computational systems used to exercise power over us, determining what we may know, what we may have, and what our options will be.

Automatic Authorities are used by both state and non-state actors.

Widespread deployment has many risks; including that we are basing increasingly important decisions on systems whose operation cannot be adequately explained to democratic citizens.

Area specialists generally agree that explanations matter. But they don't have a clear account of why.

Philosophers have only recently (e.g. Vredenburgh 2019) begun to think about the value of explanations (previously focusing on mutual *justification*). Political philosophers think too little about non-state power.

My Thesis

With rare exceptions, only if the powerful can adequately explain their decisions to those on whose behalf or by whose licence they act, can they exercise power legitimately and with proper authority, and so overcome presumptive objections to their exercise of power grounded in freedom, equality, and collective self-determination.

2. Explanations and Explainability

Explanations

Not attempting to engage with philosophy of science literature on the nature and epistemic value of explanations (e.g. Lombrozo 2011). Adopting instead a working definition focused on understanding the normative explainability debate.

To explain X is to communicate information about X that enables a presumed audience to reach a well-justified understanding of X (cp Wilkenfeld 2014).

My 'X'= acts/decisions. E.g. were you given credit? At what interest rate? Was your social media post removed or promoted? Were you granted a visa? Have you received healthcare or welfare? Have you been told to self-quarantine? Etc.

Acts can be explained by describing their causes and causal preconditions, as well as the beliefs, desires, intentions of the agent, and the processes that they followed, etc. (Malle 2004).

But to reach a **well-justified understanding of an act means different things in different contexts, and for different audiences**. That's why it's so important to know why explanations matter and to whom they are owed—this conditions what counts as a good explanation.

Explanations enable you to understand an act. **Justifications** enable you to understand the deontic status of an act. **Justifying explanations** simultaneously do both.

Explainable AI

Automatic Authorities are often **secret**, **highly complex**, and intrinsically **inscrutable** (Burrell 2016; Selbst and Barocas 2018).

E.g. recidivism prediction algorithms, DNA matching algorithms, used in courts but proprietary.

Some Automatic Authorities are very simple (e.g. Robodebt, UK grading 'algorithm'). But they are often highly complex, even when rules-based, and require high levels of expertise to understand.

When grounded in Machine Learning (ML), especially deep learning, neural nets, unsupervised or reinforcement learning, Automatic Authorities can be intrinsically inscrutable: we know they work by their results; we don't understand in any detail why or how they reach those results. Radically empiricist (Wheeler 2017). Generate unexpected correlations, and complex, counterintuitive models.

3. Power

The Nature of Power

For our purposes: power over, not power to (for discussion of definitions of power, see Dowding 2012).

A has power over B if and only if A is able unilaterally and without meaningful retaliation to make decisions that directly or indirectly affect B.

A is an agent. Ability = (roughly) sufficient probability of success conditional on trying.

Direct effects: harm or benefit.

Indirect effects: mediated by B's choice. E.g. subtracting from or adding to B's option set; altering B's options; adding penalties or inducements; surveilling B; nudging; affecting B's beliefs or desires (rationally or irrationally) etc.

Power measured in *degree* (how big the effects?), *scope* (over what range of choices?), *concentration* (ratio of Bs to As).

This working definition not uncontroversial! But bypassing objections to save time.

Justifying Power

Power of some over others is in presumptive tension with **individual freedom**, **social equality**, **collective self-determination**. But also necessary to

realise them.

A's power over B presumptively limits B's freedom, on any analysis of freedom negative, positive, republican etc. (e.g. Pettit 2008).

A's power over B places A **over** B. Presumptively undermines social equality (Anderson 1999; Kolodny 2014).

A's power over *the Bs* presumptively entails that the Bs lack power over their own destiny. Collective self-determination is to social equality (roughly) as positive liberty is to negative liberty.

These presumptive objections can be **overridden** (but residually present) when power is used to achieve worthy enough ends. They are **undercut/silenced** when power is used **legitimately**, by those with **proper authority** to do so.

4. Legitimacy, Authority, and Explanation

Legitimacy and Authority

Is power being exercised **in the appropriate way** (legitimacy), and **by the appropriate people** (authority)? E.g. Facebook Oversight Board: legitimate but lacks authority. Australia's approach to immigration/asylum: has proper authority, but lacks legitimacy.

I'm focusing on **procedural legitimacy**. Legitimacy is fundamentally about **limiting power**. Limitation in scope and degree. Mandating procedures, e.g. applying comprehensible, publicly-known rules consistently, without distinction based on morally irrelevant features. Due process. Accountability. Contestability.

These limitations serve **freedom**—reducing incursions into individual freedom and increasing security against risk of those incursions.

They serve **equality**: substantively ensuring that like cases are treated alike; also by ensuring that we (collectively) have power over those who (individually) have power over us.

This also serves **collective self-determination**, since it means we can, collectively, determine how power will be used.

Some think power is justified when it is used wisely. Some think it must also be used in the right way. I think (in addition) it must be used by the right people.

The right to exercise power derives genealogically from our right to govern ourselves. The powerful act with proper authority only when they are authorised to do so by the sovereign. The ultimate sovereign is we the people. Authorisation comes in two forms: we authorise you to exercise power **on our behalf** (= authorisation); or we authorise you to exercise power **by our leave** (= licence).

Acting on others' behalf imposes additional obligations relative to acting by their leave. E.g. you are acting with their stuff/**in their name**.

Social equality and collective self-determination require power to be exercised by those with proper authority to do so. Only then are we equals; only then are we collectively shaping the shared terms of our social existence.

Explanations

Only if the powerful are able to explain their decisions can they exercise power legitimately and with proper authority.

Explanations can reveal if a decision was intra vires—was the Automatic Authority applying the right rules?

Explanations **expose whether like cases were treated alike**, by highlighting the factors that contributed to the decision, and enabling such comparisons. They can **expose disparate treatment and disparate impact** (consider 'business necessity' defence).

Explanations **surface evidentiary bases**, so we can see whether illegitimate evidence was appealed to.

Explanations protect us against **risk of bad decisions**: not just about this case, but about whether reliable process was followed which would have given right outcome in range of other cases. Lack of explanations *necessarily* subjects us to risk.

Explanations necessary for accountability and contestability: they **show who made the decision and how they reasoned**. They are constitutive of *answerability*, the very sense of being obligated to account for one's actions (consider the phrase: 'I don't have to explain myself to you').

Explanations surface **chains of authorisation**, establishing genealogy. Again, acting within the bounds of the authorisation. E.g. who wrote these rules? Were they formed by those with proper authority to do so?

Importantly also about **acting for the right reasons**: when acting in a society's name, especially important not to elevate some perspectives over others, or to act for widely despised reasons. Explanations necessary to satisfy requirements of **public reason**.

Explanations necessary so that authorised power actually contributes to collective self-determination. **For us to conscientiously shape our social world through our proxy agents, we must understand how they are deliberating, and why they do the things they do**. Compare an astrology obsessive, who lives and dies by his horoscope. Devotedly observing gnomic guidance the basis of which one cannot understand = voluntary servitude, not autonomous self-determination.

Really important take-home: **explanations are owed, primarily, to we the people** (not in the first instance to those subject to the decision).

All of these points are true for all authorities, not just automatic ones. So let's apply this to that case.

5. Explaining Automatic Authorities

We Exercise Power with Automatic Authorities

Automatic Authorities perhaps don't exercise power themselves. But they are used to exercise power.

They **increase the degree, scope, and concentration of power**.

E.g. **automating application of penalties or inducements** through smart contracts, thereby increasing their deliberative weight. Also enabling **perpetual surveillance**: Automatic Authorities enable us to operationalise mass data collection (e.g. Office365 'compliance' tools using NLP to detect abuse; or facial recognition operationalising CCTV).

Automatic Authorities constantly shape our options, eliminating some or adding others; e.g. using **dark patterns to nudge towards giving up personal data**; delivering dynamic user interfaces shaped for us individually. They also have tremendous influence over our desires and beliefs, through **recommender systems and search algorithms** that dominate our practices of inquiry.

Automatic Authorities make it easier to exercise power, this means the powerful can influence more choices per person subject to them (increasing scope), and also can influence *more people* (increasing concentration).

Public Automatic Authorities

Roughly: public power is exercised on behalf of the whole political community; private power is exercised by or on behalf of some specific individuals or groups.

Public Automatic Authorities lack legitimacy and proper authority if they cannot give adequate explanations for their decisions.

Has admissible data only been used? Compare ImageNet (Birhane and Prabhu 2021), Clearview-AI.

Are like cases being treated alike (as far as possible within ML architecture). Is individual or structural discrimination taking place? **What alternative models would have given equally accurate results**?

Wrong kinds of reasons? E.g. **deciding based on correlation when causation is necessary**?

Lack of robustness? Protection against risk of bad decisions? **Would small perturbations in the data change the result**?

Automated systems hide accountability and authorisation—complex systems with many hands. Explanations mitigate.

Deploying ML systems in particular involves *many* engineering choices—**selecting among equally accurate models**; **tweaking hyperparameters**; **making imprecise laws precise**. Aggregate outcomes might be similar; impacts on individuals might be very different. **Who is making these choices**? Do they have authority to do so? Compare automated enforcement of copyright law.

Not just about implementation, also about *what you're optimising for*. Public reasons requirement important here. Same true for e.g. how the training data was labelled. **Many evaluative decisions are buried if we focus only on outcomes**… E.g. COMPAS *were* trying to act fairly. But **why should their CEO/engineers be the ones to decide which conception of fairness is relevant/operational**?

Explanations necessary to protect against function creep, and AI solutionism (e.g. recidivism prediction algorithms being used in sentencing).

Private Automatic Authorities

Everything I've said about public Automatic Authorities goes for private Authorities too. The question is: do the requirements of legitimacy and authority apply to them too? Some might think not…

Does our consent to use these digital platforms legitimate them? No; it's **junk consent**, with **massive externalities** for those who don't consent (see e.g. Barocas and Nissenbaum 2014). As bad an argument here as it is for the state.

Does competition void these requirements? No, and it's unlikely to because of well-known **network effects** (E.g. Barwise and Watkins 2018).

Does regulation void these requirements? Not such much void them as implement them, when done well. But regulation is not without its problems. **Putting all this new power in the hands of the state (or supra-state entities) may not enhance legitimacy**.

Do lower stakes make the difference? No, because sometimes the stakes are *very* high (e.g. Myanmar, US elections, Capitol insurrection, COVID-19). But also individually small stakes aggregate into enormous concentrations of power.

Are private Automatic Authorities really any different from other illegitimate concentrations of power? Why focus on them, rather than e.g. Murdoch, big pharma, oil etc?

Key point. Algorithmic Automatic Authorities have created a new kind of power. **As we reproduce existing social structures in digital form, we *transduce them*, changing them** in the re-presentation (Bucher 2018). Digital platforms do so in their own interest, and sometimes with the best intentions. True for markets, speech, culture, social relationships (Birch 2020).

Changing social structures is incredibly hard. **Algorithmic tools offer extraordinary promise for social progress**. They're also **unavoidable**, because there's no way to navigate the 'infoglut' of the internet without them, and because they *can* make all forms of communication and commerce more efficient (Andrejevic 2013).

The difference between this kind of power and the power of Murdoch, big pharma, big oil etc., is that their power is straightforwardly illegitimate. In a just society, Murdoch & co wouldn't exist.

But **in a just society, we *would* have institutions performing the functions performed by private Automatic Authorities**. They exercise *necessary* power. I'm tempted to call this ***governance* power—making, implementing and enforcing the constitutive rules of a socially valuable institution**.

That's why we shouldn't just aim to abolish private Automatic Authorities. We should instead aim to have them exercise power legitimately and with proper authority.

There are some differences from public Automatic Authorities. There may be good reasons for continuing to operate 'by licence' rather than 'in our name' (**some separation of powers is valuable**). This difference changes the requirements of public reason. The stakes do matter (for both public and private Automatic Authorities) where e.g. due process is concerned. We still need some measure of scrutability, because we don't want to leave the reshaping of our social structures to forces we do not understand.

6. Upshots

Highlighting these key upshots:

1. We're **definitely talking about explanations here, not justifications**. It's not about understanding the deontic status of the decision. It's about understanding how and why that decision was taken, and by whom.

2. The proper audience of an explanation is obviously often the person affected by the decision. But also **often explanations are owed to we the people, on whose behalf or by whose licence the decision was taken**. This means that some of the objections to explanations (e.g. their tendency to undermine privacy, or to create moral hazards) can be addressed through the procedures of representative democracy.

3. If the goal of explanation is to legitimate the exercise of power, then the explanation will include much more than just technical details of how the algorithm works. But those **technical details will also be important**—we need to know e.g. what data it was trained on, what modelling choices were made along the way, what values were incorporated, whether the resulting model is identifying the right kinds of relationships among variables, and so on (n.b. this is about much more than can be provided by the counterfactual approach to explanation that some favour).

References

Anderson, Elizabeth (1999), 'What Is the Point of Equality', *Ethics,* 109 (2), 287-337.

Andrejevic, Mark (2013), *Infoglut : How Too Much Information Is Changing the Way We Think and Know* (New York: Routledge).

Barocas, Solon and Nissenbaum, Helen (2014), 'Big Data's End Run around Anonymity and Consent', in Julia Lane, et al. (eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (New York: Cambridge University Press), 44-75.

Barwise, Patrick and Watkins, Leo (2018), 'The Evolution of Digital Dominance', in Martin Moore and Damian Tambini (eds.), *Digital Dominance* (New York: Oxford University Press), 21-49.

Birch, Kean (2020), 'Automated Neoliberalism? The Digital Organisation of Markets in Technoscientific Capitalism', *New Formations,* 100, 10-27.

Birhane, Abeba and Prabhu, Vinay Uday (2021), 'Large Image Datasets: A Pyrrhic Win for Computer Vision?', *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1537-1547.

Bucher, Taina (2018), *If...Then : Algorithmic Power and Politics* (New York: Oxford University Press).

Burrell, Jenna (2016), 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms', *Big Data & Society,* 3 (1), 2053951715622512.

Dowding, Keith (2012), 'Why Should We Care About the Definition of Power?', *Journal of Political Power,* 5 (1), 119-135.

Kolodny, Niko (2014), 'Rule over None: Social Equality and Justification of Democracy', *Philosophy & Public Affairs,* 42 (4), 287-336.

Lombrozo, Tania (2011), 'The Instrumental Value of Explanations', *Philosophy Compass,* 6 (8), 539-551.

Malle, Bertram F. (2004), *How the Mind Explains Behaviour: Folks Explanations, Meaning, and Social Interaction* (Cambridge, MA: MIT Press).

Pettit, Philip (2008), 'Dahl's Power and Republican Freedom', *Journal of Power,* 1, 67-74.

Selbst, Andrew D. and Barocas, Solon (2018), 'The Intuitive Appeal of Explainable Machines', *Fordham Law Review,* 87, 1085-1139.

Vredenburgh, Kate (2019), 'The Right to Explanation of Automated Decision-Making'.

Wheeler, Gregory (2017), 'Machine Epistemology and Big Data', in Lee McIntyre and Alex Rosenberg (eds.), *The Routledge Companion to Philosophy of Social Science* (New York: Routledge), 321-329.

Wilkenfeld, Daniel A. (2014), 'Functional Explaining: A New Approach to the Philosophy of Explanation', *Synthese,* 191 (14), 3367-3391.

For a copy of the full paper, please email me at seth.lazar@anu.edu.au! Many thanks for your attention.