Algorithmic Fairness and Feasibility

Eva Erman, Markus Furendal and Niklas Möller

Forthcoming in *Philosophy & Technology*

The "impossibility results" in algorithmic fairness suggest that a predictive model cannot fully meet two common fairness criteria – *Sufficiency* and *Separation* – except under extraordinary circumstances (Kleinberg et al 2017). Specifically, the results demonstrate that both fairness criteria cannot be jointly satisfied whenever the two background conditions of *Imperfect Prediction* and *Base Rate Inequality* obtain. These findings have sparked a discussion on fairness in algorithms, prompting debates over whether predictive models can avoid unfair discrimination based on protected attributes, such as ethnicity or gender.

Otto Sahlgren (forthcoming) argues that the discussion of the impossibility results would gain by importing some of the tools developed in the philosophical literature on the concept of feasibility in recent years. Utilizing these tools, Sahlgren sketches a cautiously optimistic view of how joint satisfaction of the fairness criteria can be made feasible in restricted local decision-making settings through collective action eliminating relevant base rate inequalities gradually over time.

In this paper, we argue that feasibility plays an *important* but *limited* role for algorithmic fairness. Below, we summarize the 'impossibility results' in algorithmic fairness and how Sahlgren approaches them (I). Thereafter, we analyze in what ways feasibility considerations play an important role for algorithmic fairness (III) and in what ways they play a limited role (IV). The forward-looking final section offers a sketch of a framework that may be useful for theorizing feasibility in algorithmic fairness (V).

I. Making algorithmic fairness feasible

AI systems have been criticized for making predictions based on attributes that we think should be irrelevant in the context at hand, such as when the predicted likelihood of reoffending if given parole depends on the ethnicity of the defendant. In order to avoid problematic outcomes, two independently plausible fairness criteria have been

suggested. *Sufficiency* demands that the actual outcome (whether the defendant does in fact reoffend) should be independent of the protected attribute (the ethnicity of the offender) given the outcome predicted by the system (the likelihood of reoffending). *Separation*, on the other hand, demands that the outcome predicted by the system should be independent of the protected attribute given the actual outcome.

However, it can be formally demonstrated that it is impossible to simultaneously satisfy these two fairness criteria when two background conditions obtain: *Imperfect Prediction*, meaning that there is at least one token prediction which differs from the actual outcome; and *Base Rate Inequality*, which means that the rates of the target attribute (e.g. rate of defendants reoffending) vary between groups, for the protected attribute. Since Imperfect Prediction and Base Rate Inequality seem to obtain for most realistic applications, many theorists take the impossibility results to be an unattainable ideal (e.g. Friedler et al 2016; Berk et al 2021).

Sahlgren's contribution to the debate on algorithmic fairness is to import the notion of *feasibility* from contemporary political philosophy. Feasibility, as used in this discipline, demands that norms and principles should not merely *possible* to implement, but their possibility of their implementation should not be too far-fetched; they should be, in some relevant sense, plausible to realize (Brennan and Pettit 2005; Brennan and Southwood 2007; Cohen 2009: Gilabert and Lawford-Smith 2012; Southwood and Wiens 2016).

While the precise analysis of feasibility is controversial, Sahlgren mainly utilizes the rather uncontroversial distinction between hard and soft feasibility constraints. Hard constraints are those that cannot be lifted, such as metaphysical, logical and nomological constraints. Soft constraints, by contrast, are malleable and thus subject to dynamic variation, such as cultural, institutional, economic, psychological and motivational constraints. Rather than strictly speaking excluding a state of affairs, they make various possible states of affairs more or less likely to be realized (Gilabert and Lawford-Smith 2012). For example, if the ideal of a fair distribution of resources requires the state to increase income and capital tax, such a policy might be undermined by disincentive effects and tax evasion. When assessing a theory, it matters whether these are best understood as hard or soft constraints.

Sahlgren utilizes the observation that while the background condition of Base Rate Inequality arguably will obtain in most realistic cases, there is actually no hard constraint

prohibiting us from attaining base rate *equality*.[1] Instead, there are various soft constraints that make base rate equality more or less likely. Sahlgren investigates contexts in which it is more feasible, such as aiming for more locally contextualized predictive models than the current preference for general – perhaps even universal – ones. Within a restricted population, Sahlgren argues, reaching (or coming close to) base rate equality becomes more feasible.

A second obstacle, Sahlgren suggests, is the predominant focus on what is 'synchronically feasible', feasible *here and now*, which hides the equally important question of what is 'diachronically feasibility', feasible at a later point in time. Sahlgren suggests several considerations which make it plausible that given time and the right set of actions –technological and political – we may reach a change in the social system generating the data sets such that what is now (synchronically) infeasible becomes (diachronically) feasible.

II. The importance of feasibility in algorithmic fairness

The main contribution of Sahlgren's article is that it demonstrates the significance of importing insights from the feasibility debate into the debate on algorithmic fairness. Since AI systems directly impact people's lives, it is important that theories of algorithmic fairness can realistically be implemented. Impractical fairness frameworks may lead to systems that are normatively inconsistent, unpredictable, or even harmful. Feasibility considerations help ensure that fairness translates effectively into real-world applications, balancing normative ideals with technical and other empirical limitations.

There are several interconnected ways in which feasibility considerations may shape the development and deployment of fair algorithms generally. As exemplified by the impossibility results, different fairness criteria often cannot be satisfied simultaneously and recognizing this feasibility constraint is key as it prevents developers from pursuing impractical combinations of fairness goals that cannot coexist. Since feasibility is context-sensitive, it encourages prioritizing fairness based on specific use cases. Moreover, different domains – such as criminal justice, healthcare and hiring – typically require different fairness considerations and feasibility ensures that these

[1] The first background condition, *Imperfect Prediction*, is not discussed by Sahlgren; presumably because perfect prediction will most likely be unattainable for virtually any relevant application.

considerations are not overly rigid but instead adaptable to various contextual circumstances, allowing them to provide ethical guidance.

This context sensitivity, however, raises questions about the scope of Sahlgren's contribution. What are the implications of showing (or rather plausibly suggesting) that base rate equality is feasible to obtain or approximate in some specific contexts? Indeed, we can typically fulfil even highly ideal normative criteria in political theory if the domain of applicability is limited enough. So, the question is how useful this insight is for algorithmic fairness. This also illustrates a more general point, beyond Sahlgren's contribution: an important task for research on algorithmic fairness is to answer the question of when it matters whether algorithmic decision-making lives up to criteria like these. We suggest that this might be more important in certain forms of decision-making than others, and that it ultimately depends on which forms of algorithmic unfairness matter. For instance, when does it even matter whether base rate equality is obtained? Decisions that mirror or reinforce unfairness between men and women, for example, seem salient, but for some other groups – such as people with different shoe sizes – perhaps not. Here, the debate on algorithmic fairness can draw on valuable discussions around fairness and discrimination more generally (Lippert-Rasmussen 2024). We may also ask: what conclusions should be draw if base rate equality is not obtained? Should we then refrain from using AI in decision-making? What is our second-best option? As we will discuss next, in responding to these questions, theories of feasibility are of limited use, since such responses require normative analysis. While the concept of feasibility is crucial for couching normative theories in practical terms, as a way to guide real-world action and governance, it has its limitations.

III. The limitations of feasibility in algorithmic fairness

Generally, feasibility considerations are important in order to draw the best conclusions about what algorithmic fairness requires. But while the task of investigating whether algorithmic fairness conflicts with hard constraints like logical or natural science laws is relatively straightforward in principle, things become more complicated with soft constraints, such as economic limitations, social and political norms, and organizational resistance. Because as soon as we move from hard to soft constraints, algorithmic fairness becomes a *gradual* rather than a binary affair. Indeed, even computational limits – i.e.,

that algorithms have finite processing power and storage, which may limit the complexity of fairness models they can execute – which intuitively may seem 'hard', are typically not a hard feasibility constraint, given the rapid speed of computational development.

In what sense does this mean that feasibility considerations are of limited use for theories of algorithmic fairness? First, there is an intense debate in political philosophy between so-called ideal and non-ideal theory over how much weight should be given to feasibility constraints at all in designing and justifying normative theories. One lesson from ideal theory is that we should always be suspicious of arguments about constraining what we ought to do on what we take to be more or less plausible to bring about. Soft feasibility constraints are by definition contingent and subject to change, and what has been perceived as 'infeasible', indeed even 'impossible', at one point in time (e.g. abolishing slavery, giving voting rights to women) have turned out to be evidently feasible at a later point in time. Hence, the problem with restricting ourselves to only suggesting 'feasible norms' is that it may hamper our suggested normative principles in a too conservative way. Downplaying the relevance of soft constraints may hence sometimes offer a useful moral compass, providing aspirational goals and pointing out in which direction we should go (at least eventually).

Second, the gradual nature of soft constraints means that what we ought to do in relation to them is always an open question. Sahlgren is right in his observation that at least typically, the more local a context of application is, the more likely it is that we can achieve both Sufficiency and Separation. But from a fairness point of view, we might prefer some degree of base rate inequality in order to have a larger context of application. The overall fairness of society, for example, may benefit even when we cannot satisfy both fairness criteria. Likewise, it might be that we prefer slightly less algorithmic fairness in the short-term (Sahlgren's synchronic aspect) if it produces more fairness in the long term (diachronic aspect). Such questions can only be answered through a process of normative analysis for the case at hand.

IV. A sketch of a framework

Incorporating feasibility considerations into algorithmic fairness could be said to correspond to a move from 'ought implies can', i.e. a possibility constraint on normative theories, to 'ought implies feasible': only if it is feasible to follow a set of normative

demands, do we have a duty to do so.[2] Provided we do not set the bar for feasibility too low (such that, for example, a state of affairs is feasible only if it is very easy to obtain), it seems like a reasonable constraint on principles of algorithmic fairness, given their applied nature. So far in this paper, we have addressed the importance of feasibility considerations as well as their limits. We will end the paper by briefly sketching a framework consisting of two metatheoretical constraints on normative political principles, through which we elucidate several central aspects for determining the proper feasibility considerations to be made when theorizing principles of justice and fairness in the AI domain. Indeed, these constraints are of particular importance in the kind of contextual applications of algorithmic fairness that Sahlgren has in mind.

As we have stressed, feasibility considerations put the normative complexity in focus. How should we navigate between different ideas of justice and fairness, when some states of affairs are clearly feasible to bring about, but are perhaps less normatively desirable than other, less feasible ones? This is a substantial question for applied normative theory to deal with. Sahlgren's point that moving from predictive models that aim to have a general, perhaps even universal application in a certain area – such as predicting recidivism or economic viability – to a more local context to turn previously infeasible background conditions into feasible ones, echoes a broader contextual trend in how feasibility is theorized in philosophy (Erman and Möller 2020; Southwood 20202). In a recent paper, we call this the *functional constraint* (Erman and Möller 2020). In its most abstract form, it says that the guiding principles of a normative account must be appropriate for what the account aims to do, that is, what the suggested principles are supposed to regulate, and within what limits they are supposed to do so. Normative theories have a multitude of aims. According to some theorists, what we morally ought to do in a situation is the question of what we should do, all things considered (see, e.g. Hare 1952; Gibbard 2003). In political theory, however, the output of a normative account is typically not a principle which aims to determine what we should do, all things considered. Rather, more limited questions are asked, such as what is 'demanded by justice' or 'what is fair'. The functional constraint asks us to specify – or at least explicitly

[2] See Southwood 2016: 9; Southwood and Wiens 2016: 3043, and Brennan and Southwood 2007 for versions of this condition.

consider – the limits of our normative endeavor. More specifically, it emphasizes three aspects as vital when construing normative theories.

One aspect is what kind of normative *principles* we aim for. If the sought principle is a principle of justice, it arguably cannot be a principle which suggests that all goods should be distributed to the persons born in 1983, or something similarly arbitrary. We think this aspect is undertheorized in the debate on algorithmic fairness. While the principles explicitly discussed in this debate are labelled 'fairness principles', the domain is actually broader: we want our predictions to be *just*, and that arguably involves further considerations than mere fairness – in particular if conceptualized as the two criteria (Sufficiency and Separation) discussed in this debate. For different decision situations and application contexts, different considerations come into play. In one context where equal base rate is infeasible to achieve, a predictive model may still be the most just to implement, given the available decision procedures. Only a case-by-case normative analysis may settle the case.

A closely related aspect is what kind of *practice* our predictive models are to be applied to. Take a system for job application selection. Here, differences in practices arguably motivate different principles of selection, and therefore different attitudes towards certain properties. For some positions, such as the appointment of professors at a university, it seems reasonable to utilize the Rawlsian idea of positions open to all, with meritocratic concerns playing a major role. For other positions, say, the just appointment of a leader of a family business, meritocracy may perhaps justifiably play a less deciding role.

A third aspect of the functional constraint is the *temporal* aspect. Indeed, whether or not a principle is feasible depends in large parts on when the principle is supposed to come into effect. A state of affairs that is infeasible to bring about today might be feasible tomorrow, and vice versa. This means that there is a multitude of potential feasibility considerations depending not only on where, but also on *when* our predictive models are to be implemented. While this is an explicit consideration in what Sahlgren calls 'diachronic feasibility', his synchronic-diachronic distinction overlooks the non-binary nature of this consideration. 'Diachronic' might here refer to both next year's data set, those resulting from a far-away future, and everything in-between. The closer in time we suppose our principles to apply, the less deviation from the current, non-ideal reality we can expect. Again, this is an aspect that should figure in our normative analysis.

Apart from the functional constraint, a second constraint which also comes into play in theorizing principles of justice and fairness, is what we call the *fitness constraint*. The fitness constraint captures the idea that a normative principle of an account should fit together with the other principles, values and states of affairs which are endorsed in the account. If there appears to be tensions between different commitments of the account, regardless of other virtues, these must be resolved before the account may be considered justified. And the resolution may only be made in either of two fundamental ways: by abandoning at least one of the commitments in the account, or by showing that there actually is no tension after all (see Erman and Möller 2018).

Although the fitness constraint is in accordance with the method of reflective equilibrium (Rawls 1951, 1999), we do not think it is correctly described as a methodological constraint. Rather, it is compatible with virtually any methodology. The fitness constraint puts a dynamic condition on the principles that the account sets out to justify. Any direction of justification, whether bottom up or top down, or a set of commitments on an equal level of prior justificatory force, is allowed. This is the case because the conditions of fitness concern the ways in which all relevant claims fit together in the account. Simply put, in order for the account to be justified, it must fit with the other claims on which the account is premised. The fitness constraint, as we see it, is yet another emphasis on the non-static and contextual nature of normative analysis that we cannot avoid if we are to ensure, for example, that our AI systems make as just predictions and assessments as possible.

This framework broadens the role of feasibility. Applying it to algorithmic fairness, it becomes clear that feasibility considerations come into play in many places in our theorizing. Apart from being central when thinking about how to fulfil a specific background condition (base rate equality) for specific principles of algorithmic fairness (the criteria of Sufficiency and Separation), as shown by Sahlgren, it is perhaps even more important for selecting which principles are most justified in the first place, given what they are supposed to regulate and within what limits they are supposed to do so. To achieve this, however, requires that we broaden our analysis beyond the narrow context of applicability of a predictive model, to include the broader political and institutional context. Indeed, the latter is a challenge for the debate on feasibility more generally. With few exceptions, philosophical approaches tend to collapse political feasibility into the feasibility of individuals' aggregated political actions (for exceptions, see Gilabert and

Lawford-Smith 2012). Yet, in contrast to individual action, politics is an essentially collective enterprise which takes place in an institutional environment characterized by conflicting interests, ideological disagreement, imperfect knowledge, and uncertainty about the long-term consequences of specific regulatory frameworks. Hence, even in cases where algorithmic fairness is technically and economically feasible, it is typically politically challenging.

References

Beigang, Fabian. 2023. "Reconciling Algorithmic Fairness Criteria." *Philosophy & Public Affairs* 51(2).

Berk, Richard, et al. 2021. "Fairness in criminal justice risk assessments: The state of the art." *Sociological Methods & Research* 50(1): 3-44.

Brennan, Geoffrey and Nicholas Southwood. 2007. "Feasibility in Action and Attitude." In *Hommage á Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, ed. T. Rønnow-Rasmussen *et al*. Lund: Philosophy Department at Lund University.

Brennan, Geoffrey and Philip Pettit. 2005. "The Feasibility Issue." In *Oxford Handbook to Contemporary Philosophy,* ed. F. Jackson and M. Smith, 258–297. Oxford: Oxford University Press.

Cohen, G. A. 2009. *Why Not Socialism?* Princeton, NJ: Princeton University Press.

Erman, Eva and Niklas Möller. 2018. *The Practical Turn in Political Theory*. Edinburgh: Edinburgh University Press.

Erman, Eva and Niklas Möller. 2020. "A World of Possibilities: The Place of Feasibility in Political Theory." *Res Publica* 26: 1-23.

Friedler, Sorelle, Carlos Scheidegger and Suresh Venkatasubramanian. 2016. "On the (im)possibility of fairness." arXiv:1609.07236.

Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Gilabert, Pablo and Holly Lawford-Smith. 2012. Political Feasibility: A Conceptual Exploration. *Political Studies* 60: 809–825.

Grant, David Gray. 2023. Equalized odds is a requirement of algorithmic fairness. *Synthese* 201(3).

Hare, R. M. 1952. *The Language of Morals*. Oxford: Oxford University Press.

Hedden, Brian. 2021. "On statistical criteria of algorithmic fairness." *Philosophy & Public Affairs* 49(2): 209-231.

Hellman, Deborah. 2020. "Measuring algorithmic fairness." *Virginia Law Review* 106(4): 811-866.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. "Inherent trade-offs in the fair determination of risk scores." In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (ITCS'17).

Lippert-Rasmussen, Kasper. 2024. "Algorithmic and Non-Algorithmic Fairness: Should We Revise Our View of the Latter Given Our View of the Former?" *Law and Philosophy*, https://doi.org/10.1007/s10982-024-09505-4.

Long, Robert. 2021. "Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness." *Journal of Moral Philosophy* 19(1): 49-78.

Rawls, John. 1951. "Outline of a Decision Procedure for Ethics." *Philosophical Review* 60: 177–197.

Rawls, John. 1999. *A Theory of Justice*. 2nd eds.. Cambridge, MA: The Belknap Press of Harvard University Press.

Sahlgren, Otto. Forthcoming. "What's Impossible about Algorithmic Fairness?" *Philosophy & Technology*.

Southwood, Nicholas, and David Wiens. 2016. "Actual' Does Not Imply 'Feasible". *Philosophical Studies* 173: 3037–3060.

Southwood, Nicholas. 2022. "Feasibility as Deliberation-Worthiness", *Philosophy & Public Affairs* 50 (1): 121-162.