The Unsatisfiable Triad: A Problem for Automated Decision Making

Sebastian Sequoiah-Grayson
School of Computer Science and Engineering
University of New South WalesSydneyAustralia

Abstract

In this paper, we argue that the properties of *fairness*, *accountability*, and *transparency* form a consistent, satisfiable triad for Human Decision Making (HDM). We then argue that these same properties form an *unsatisfiable* triad for Automated Decision Making (ADM). That is, for ADM, the entrenchment of one of these properties will make at least one of the others more remote.

We state the rationale for the consistent triad for HDM via a belief/desire psychological model. (1) We identify decisions with actions of a psychological sort. (2) We identify *accountability* with the provision of an explanation. (3) We adopt a causal model of explanation. (4) We specify causal relations via counterfactuals. (5) We identify beliefs and desires as the independently necessary and jointly sufficient psychological attitudes providing the relevant counterfactual support. (6) We argue that this explication of accountability facilitates a consistent, satisfiable triad with respect to *fairness* and *transparency*. (7) We make the claim for the latter via appeals to testimony and declaration. (8) We make the claim for the former via appeals to conjunctions of beliefs and desires being sufficient not only for the specification of the aforementioned causal antecedents, but for the specification of reasons with normative salience. (9) We conclude that this normative salience is plausibly sufficient for fairness with regard to the relevant target decision for which we aim to provide an account in the first instance.

We then argue that this consistent triad is unavailable for ADM, and that any obvious attempt to pursue it reinforces the ADM triad's unsatisfiability. We start by introducing two ADM use cases - *recidivism prediction* and *LLM detection*. (10) We argue that *accuracy* is a necessary but insufficient condition on *fairness* for both use cases. (11) We argue - in line with (1)-(8) above - that *transparency* is a conjoint necessary condition on fairness for decision making. (12) We argue that accuracy in ADM turns on computational complexity, and that *computational complexity inhibits transparency*. (13) Via (10) and (12) we then conclude that for ADM, there is an inverse-relationship between fairness and transparency. (14) Via (7), we argue for a corollary of (13), that for ADM there exists an inverse relationship fairness and accountability.

We explore attempts to repair the triad for ADM by replacing the causal model of explanation with a statistical model of explanation. We do not find this promising. We conclude that the unsatisfiability of the *fairness*, *accountability*, and *transparency* triad for ADM remains an open problem. We recommend that for those decisions for which the triad matters - and we believe that these are many - that humans must remain in the decision-loop.

1 Introduction

We argue below that the properties of *fairness*, *accountability*, and *transparency* form a consistent, satisfiable triad for Human Decision Making (HDM). We then argue that these same properties form an *unsatisfiable* triad for Automated Decision Making (ADM). That is, for ADM, the entrenchment of one of these properties will make at least one of the others more remote.

We state the rationale for the consistent triad for HDM via a belief/desire psychological model. (1) We identify decisions with actions of a psychological sort. (2) We identify *accountability* with the provision of an explanation. (3) We adopt a causal model of explanation. (4) We specify causal relations via counterfactuals. (5) We identify beliefs and desires as the independently necessary and jointly sufficient psychological attitudes providing the relevant counterfactual support. (6) We argue that this explication of accountability facilitates a consistent, satisfiable triad with respect to *fairness* and *transparency*. (7) We make the claim for the latter via appeals to testimony and declaration. (8) We make the claim for the former via appeals to conjunctions of beliefs and desires being sufficient not only for the specification of the aforementioned causal antecedents, but for the specification of reasons with normative salience. (9) We conclude that this normative salience is plausibly sufficient for fairness with regard to the relevant target decision for which we aim to provide an account in the first instance.

We then argue that this consistent triad is unavailable for ADM, and that any obvious attempt to pursue it reinforces the ADM triad's unsatisfiability. We start by introducing two ADM use cases - *recidivism prediction* and *LLM detection*. (10) We argue that *accuracy* is a necessary but insufficient condition on *fairness* for both use cases. (11) We argue - in line with (1)-(8) above - that *transparency* is a conjoint necessary condition on fairness for decision making. (12) We argue that accuracy in ADM turns on computational complexity, and that *computational complexity inhibits transparency*. (13) Via (10) and (12) we then conclude that for ADM, there is an inverse-relationship between fairness and transparency. (14) Via (7), we argue for a corollary of (13), that for ADM there exists an inverse relationship fairness and accountability.

We explore attempts to repair the triad for ADM by replacing the causal model of explanation with a statistical model of explanation. We do not find this promising. We conclude that the unsatisfiability of the *fairness*, *accountability*, and *transparency* triad for ADM remains an open problem. We recommend that

for those decisions for which the triad matters - and we believe that these are many - that humans must remain in the decision-loop.

2 Human decision making

2.1 Attitudes and actions

The familiar, everyday, or *folk* understanding of human psychological attitudes has well-known explanatory virtue with regard to human behaviour [**stich1994folk**], [**fodor1996folk**], [**horgan2013folk**]. We attribute attitudes such as believing, fearing, desiring, wishing, loving, suspecting, wanting, predicting, and so on, as playing central, first-class roles when it comes to understanding human action.

For example, suppose that Alice stands from her chair and walks over to the cupboard, opens it and gets a biscuit from the tin, and then eats the biscuit. In this case we may suppose that Alice's want for a biscuit in combination with her belief that there were biscuits in the cupboard goes some way to providing an account of her biscuit-seeking behaviour. In this example, our target behaviour of Alice's is her *physical* behaviour. It need not be however. Psychological behaviour is behaviour through and through, albeit of a uniquely *mental* sort. Indeed such psychological behaviour might precipitate much of physical behaviour.

For an example of such psychological behaviour, suppose that Alice is in the process of *making a decision*. The verb-clause here is not redundant. Decisions are things that humans do, or carry out. Witness the same predicate in one-place, property-denoting form - *Alice is deciding*. Psychological attitudes will figure into accounts of psychological behaviour in much the same way as they do in accounts of physical behaviour. Suppose that Alice decides to follow the white rabbit. In this case we may suppose that Alice's desire to know where the white rabbit is going, in combination with her belief that following the white rabbit was the best course of action to pursue with regard to satisfying her aforementioned desire, goes some way to providing an account of her *deciding* to engage in her rabbit-following behaviour in the first place.

In the case of both physical and psychological behaviour, we frequently understand the psychological attitudes of the agent in question to play a constitutive role in the provision of an account of such behaviour. In order that such an account provide be a real and true account of such behaviour, it must provide an *explanation* of that same behaviour.

Note that descriptive adequacy alone is necessary, but insufficient for accountability. We will see below how it is that accountability requires *reasons* also, and how it is that explanations may figure into the provision of such reasons. We begin with explanation.

2.2 Explanation, causation, and counterfactuals

Explanation is not primitive. Familiar models of explanation include the *deductive nomological* model [**hempel1962deductive**], the *statistical model* [**salmon1971statistical**], [**niiniluoto1981statistical**], the *pragmatic model* [**van1988pragmatic**], and the *causal* model [**lipton1990contrastive**]. Although each model arguably has advantages in various contexts, we pursue a causal model of explanation here as we understand it to have the most natural fit with our phenomena of interest.

On the causal model of explanation, to provide an explanation of some event or state of affairs x is just to specify the cause or causes of x. This fits across a range of scenarios. Causal antecedents are accepted as explanations in many contexts, from medical diagnoses to accident investigations. In such contexts, it would be out of place for a putative explanation to not figure into the antecedent causal structure.

Like explanation itself however, causation has been modelled variously. From outright scepticism via *constant conjunction* models [**beebee2016hume**], to realist models of both *sufficiency* and *counterfactual* [**lewis1986causation**] varieties. We take the counterfactual model of causation to be both the most plausible, and the best fit with the phenomena at hand. The counterfactual model of causation takes causes to be *necessary* conditions on their effects. This implies that had a cause not occurred, then the relevant effect would not have eventuated.

This route from counterfactuals to explanation via causation fits our Alice-based examples of HDM above. Consider Alice's biscuit-seeking. We may understand Alice's want for a biscuit, in combination with her belief that the biscuits are in the cupboard, as an explanation for her behaviour. We may understand these psychological attitudes of Alice's as providing the explanation at hand in virtue of the fact that they *caused* Alice to leave her chair and walk to the cupboard. In combination with the absence of defeater attitudes and physical obstructions and so on, we may understand these attitudes of Alice's as being jointly sufficient for her behaviour. In virtue of these attitudes being the cause of Alice's behaviour, we may understand them as being independently necessary. That is, with all things being equal, had either of Alice's relevant attitudes not been the case, then her biscuit-seeking behaviour would not have occurred.

Similarly with the example of Alice's decision-making. We may understand Alice's desire to know where the white rabbit was going, in combination with her belief that following the white rabbit was a suitable way to satisfy her desire in question, as an explanation for her deciding to follow the white rabbit in the first place. We may understand these psychological attitudes of Alice's as providing the explanation at hand in virtue of the fact that they *caused* Alice to make the decision that she made. In combination with the absence of defeater attitudes, we may understand these attitudes of Alice's as being jointly sufficient for her psychological behaviour. In virtue of these attitudes being the cause of Alice's behaviour, we may understand them as being independently necessary. That is, with all things being equal, had either of Alice's relevant attitudes not been the

case, then Alice would have failed to decide as she did.

Noting again that accountability requires *reasons*, and having outlined how it is that explanations may be understood in terms of a causal-counterfactual model, we leverage this schema so as to see how it may figure into the provision of reasons themselves. In this way we will make the connection for HDM between accountability on the one hand, and fairness along with transparency on the other.

2.3 From accountability to fairness and transparency

The explication above of accountability in terms of explanation allows us to connect accountability with fairness. This is by appeal to psychological attitudes being sufficient not only for the specification of the aforementioned causal antecedents of human behaviour, but for the specification of reasons for that behaviour, where such reasons import normative salience.

This is not obvious. In general, explanations are not *ipso facto* justifications - witness motive vs exculpation. Similarly, causes are not reasons, at least not in any general sense that confers reasonableness in its permissive sense. In Alice's case however, the causes of her behaviours above *are* her reasons - in the full normatively salient, permissive sense. This is on account of their being the attitudes of Alice's that Alice herself might be expected to proffer should she be asked to justify, ground, provide warrant for, or provide a rational basis for, her behaviour [**kearns2008reasons**], [**brunero2013reasons**], [**ginet2005reasons**], [**barocas2020hidden**]. That is, in those cases where she is asked to demonstrate her behaviour's reasonableness.

Not all behaviour - physical or psychological, is reasonable, but it may be reasonable insofar as it is caused by attitudes of the agent of an appropriate sort. In this way, a causal explanation of behaviour may be a demonstration of that behaviour's reasonableness - in the rich permissive sense. By reporting one's attitudes as the causes of one's behaviour, one provides an account of one's behaviour that presents the behaviour as reasonable, and *fair*. We may and do say of one's reasons that they are "fair reasons" for good reason. One's act of decision making will be accountable when it is reasonable, and its reasonableness will be a function of the extent to which it is grounded, justified, warranted, or given normatively rational salience, by one's causally antecedent attitudes. In those situations where the attitudes proffered are accepted as normatively salient, we say that the decision is a *fair decision*.

The *reportability* of an agent's normatively salient attitudes as the causal antecedents of their decision-making behaviour is the property that connects accountability to *transparency*. Our reasons for making a decision - the account of our decision that we provide - are *transparent* in virtue of the fact that they are accessible by introspection and transmissible by testimony.

For HDM then, accountability, fairness, and transparency form a satisfiable triad. For Automated Decision Making (ADM), this is not the case. The entrenchment of one of these properties will serve to make at least one of the others more remote.

3 Automated Decision Making

3.1 Two ADM use-cases

3.1.1 COMPAS recidivism prediction software

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) recidivism system is an algorithm built for the purpose of predicting recidivism of convicted criminals facing sentencing in court. The goal of the COMPAS system is to assist the sentencing judge with their sentencing decision. The system's algorithm outputs a recidivism prediction score using a database of offender demographic data points. COMPAS's recidivism prediction score is then used by the presiding judge for sentencing.

3.1.2 Turnitin LLM detection software

Turnitin is plagiarism detection software used in universities worldwide. It checks students' essays against a database of previously submitted essays, as well as web-crawled repository of material published online. In 2023, Turnitin expanded their plagiarism detection software to include a function designed to detect the use of large language model based generative artificial intelligence. This is well-motivated on the face of it. Given that LLM generated content will not pre-exist online, bespoke detection algorithms are required in order to guard against obvious threats to academic integrity.

3.2 From accuracy to accountability

The COMPAS system is well-known across AI circles for its predictive shortcomings. From racial and gender bias, to outright predictive error. The system has become a near trope with regard to the risks of delegating decision-making to automated systems. Concern with regard to these risks was not mitigated by the parent company Northpointe (now Equivant) conflating *type 1* (false positive) and *type 2* (false negative) errors in their initial response [**dressel2018accuracy**], [**rudin2020age**], [**chouldechova2017fair**], [**wang2023pursuit**].

What these and related concerns have in common is that they are concerns about *accuracy*, or lack thereof. This is well-motivated. Accuracy is a necessary condition on any decision, automated or otherwise. However, in the case of ADM, it is further from the only crucial issue than is obvious.

To put this in relief, let us assume the scenario where most if not all concerns about accuracy have been mitigated, and that the accuracy of the COMPAS system with regard to recidivism prediction far exceeds that of any human agent. Now consider the following state of affairs with two human agents, one the sentencing judge, and the other the convicted criminal undergoing sentencing by the judge in question. Suppose that the exchange between the judge and the criminal goes something along the lines of the following. J: "I hereby sentence you to ten years in Goal, with a non-parole period of six years". C: "Your honour, that is a bit rough. Why is my sentence so harsh". J: "Your sentence is

severe, because I believe that there is a high likelihood that you will re-offend". C: "Your honour, why have you drawn the conclusion that I shall re-offend?". J: "I have drawn the conclusion that you will re-offend on account of the prediction of the same being made by COMPAS. I shall have you know that COMPAS is very rarely if ever wrong in these matters". C: "But your honour, why did COMPAS make such a prediction?". J: "Well you see...COMPAS has many trillions of parameters. After all, this is what makes it so accurate! Neither I, nor anyone else can be expected to actually understand *how* it is that COMPAS arrives at its predictions, but the good thing is that we can all be reassured that it gets things right". C: "Well this does not sound terribly reassuring to me at all".

We think that the criminal has a point. The scenario above (and others like it) is worryingly dystopian. This is borne out by reflecting on the triad of accountability, fairness, and transparency explored above. It is the (assumed in our scenario) accuracy of the COMPAS system that allows it to achieve fairness, but this comes at the cost of large computational complexity. So large, in fact, that it mitigates any possibility of an account - an explanation in familiar causal terms or otherwise - of how it is that COMPAS arrived at its decision. Unlike human agents, COMPAS does not possess anything like psychological attitudes. There is nothing of this sort to which we can appeal. Moreover, there is nothing happening within COMPAS situated within the space of reasons in any sense. Certainly not in any sense that could confer reasons with normative salience. COMPAS can be fair only in the sense that it can be accurate. The unaccountability of the COMPAS system for its outputs is the *cost* borne by the fairness of its decisions, insofar as such fairness is a function of its accuracy, which is in turn a function of its computational complexity.

The same computational complexity of the COMPAS system mitigates transparency also. The large number of its parameters is sufficient to make its internal machinations opaque. This is the point made by the judge in our scenario above. The situation is clear - fairness for the COMPAS system imposes a level of complexity that rules out transparency and accountability. Similarly, increasing transparency would come at the cost of computational simplification, and computational simplification would decrease accuracy, thereby decreasing the fairness of the COMPAS system.

The scenario above might be considered to be overly speculative. A more pedestrian example, and one familiar to most practicing academics, is that of Large Language Model (LLM) detection systems. One of, if not the most popular of these is that incorporated into the Turnitin plagiarism detection software package. We note that recent LLM detection software has not been terribly accurate. The author(s) has/have experimented with this first-hand, and the results have been mixed at best, although we have anecdotal evidence that they are improving. However, the author(s) is/are also practicing in academia, and is/are working with Turnitin on a regular basis in a professional setting. As such, we have real-world first-hand familiarity with a large number of scenarios that map on to the following scenario very closely.

There are three agents, a Lecturer, and Tutor, and a Student. The student

submits an assessment - an essay say - through the Turnitin system - as is obligatory at their university. The student's essay scores a high probability from the Turnitin system as having resulted from the use of an LLM/generative AI resource such as ChatGPT. The tutor, who is also the marker of the student's assessment, contacts the lecturer with obvious concerns. But what, asks the lecturer, is the tutor - or for that matter the lecturer themselves - supposed to say to the student? That a piece of proprietary software (so we have already mitigated transparency) has returned an LLM score that is unfavourable to the student? In step with the recidivism case above, what could either the lecturer or the tutor say to the student when challenged? That although they do not know why it is that the LLM detection system has returned the score that is has returned, the student is still in trouble for violating rules surrounding academic integrity because the incriminating numbers came up, and the system is rarely wrong? With the exception of the most brazen cases of academic misconduct, it is hard to see how such a scenario could get much further than a standoff.

The academic scenario above is common place, and serious [**tang2023science**]. In the modern academic environment, large sums of money, professional reputations, academic progress, and for some students visa status, are all at stake. The unsatisfiablity of the triad of accountability, fairness, and transparency is as concrete for the academic scenario as it is for that of recidivism predictions.

We do not believe that this unsatisfiability is special to recidivism prediction and LLM detection. We believe that it is an inevitable consequence for all cases of ADM as things stand. This is due to accountability and fairness being mutually exclusive for ADM due to the aforementioned facts with regard to large computational complexities.

4 Statistical explanation

A tempting alternative is to move from a causal model of explanation to a statistical model of explanation [**salmon1971statistical**], [**niiniluoto1981statistical**]. Given that generative ADM is working on the back of large sets of weighted parameters, permitting these to play a first class role in explanation *qua* weighted parameters is an obvious stance. There are two preemptive obstacles however - one epistemic, the other normative.

The epistemic obstacle is that confidence in such models turns on their rate of statistical success, which is statistical in turn. Such bootstrapping might be palatable for those of us whose confidence in the outputs of the weighted parameters is already high - in which case our search is for a rational basis for the beliefs that we hold already. For those of us whose confidence is not already high however, such bootstrapping will be question-begging, and viciously so.

The normative obstacle is this. Given that the statistical explanations are statistical through and through, it is not obvious how it is that they could be presented as reasons - in the richly normative sense that ties together the triad of accountability, fairness, and transparency. This normative obstacle cast doubt on the cogency of statistical explanation itself.

The schema for statistical explanations, what we will call *schema S* is as follows:

(1) Events of type T have probability p of occurring.

(2) e is an event of type T.

– –

(3) e occurs.

The dotted line in S marks *explanatory validity* - if (1) and (2) are true, then they explain (3).

Varieties of statistical explanation that hold that for p to have explanatory salience the value of p must be high are known as *elitist* varieties. Those that permit explanatory salience for low values of p are known as *egalitarian*.

However, no matter whether we adopt an elitist or egalitarian position on statistical explanation, we have grounds to doubt that anything of the form captured in Schema S above can provide a statistical explanation - on account of there not being any statistical explanations of particular individual events.

By way of example, consider a fair coin toss and the following instance of Schema S:

(1′) Events of type T - a fair coin landing heads on a toss after a suitably high number of tosses - have probability 0.5 of occurring.

(2′) h is an event of type T.

– –

(3′) h occurs.

To be sure, facts relevant to (1′) can and do explain why it is that the coin has a 0.5 chance of coming down heads. However, this is *an explanation of the value of* p *itself* - in this case 0.5. What it is *not* is an explanation why it is that the coin landed heads, *in this particular event* h *itself*.

To put this another way, statistical explanations are (putative) explanations of why it is that a value p is the likelihood value for some event. What statistical explanations are *not* are explanations of particular events. Whether such events are fair coins landing heads instead of tails, or a generative AI outputting one string as opposed to another is irrelevant. Irrelevant also is the value of p, hence this objection to statistical explanation holds for both elitist and egalitarian varieties alike.

Rather, if we want a genuine *explanation* of (3′) itself, then we must appeal to the relevant causal antecedents. These will include the relevant counterfactually salient facts about the various physical properties for the coin and the toss that led to the event h.

As a final aside, we might wonder if there any non-causal varieties of explanation that provide genuine explanations. This is doubtful. To generalise

slightly further than the above - grounds for expecting an outcome are never explanations - not when those grounds are divorced from causal antecedents.

To see this, consider again a coin toss, but this time with a biased/loaded coin such that it comes up heads much more often than otherwise. In this case, the coin coming up heads disproportionately after a long run of tosses gives one strong grounds for expecting that the coin is biased, but it does not explain the bias.

In light of the arguments above, and in spite of the initial plausibility of statistical models of explanation for satisfying *fairness*, *accountability*, and *transparency* jointly for ADM, statistical iis not an option.

5 conclusion

The satisfiability of the triad is a necessary condition for many of our most valued interpersonal relations and our most treasured social institutions. The apparent unsatisfiability of the triad for ADM should cause us concern, and motivate further work on this problem.

There is the response that this is not really a problem at all, and that the resulting gaps in accountability might even be morally desirable. The idea here is that accountability entails responsibility, and responsibility for difficult decisions can impose strongly negative emotions on the decisions makers. All things being equal (so this response goes), a world where there were less strongly negative emotions born by decision makers than there are otherwise would be a world to be preferred.

We disagree. The argument above conflates the subjective phenomenology of negative emotions on the one hand, with the fact that they are born for good reason on the other. Making a difficult decision - in a scenario where there are no good ones say - is *supposed to make us feel bad. That it does so* is a mark of good moral character. In line with even the crude utilitarianism motivating the argument above, we say that feeling good *about the fact that* we felt bad because of the decision we made is a mark of moral fortitude. If this is correct, then arguments purposed towards the mitigation of such fortitude are the mark of no good thing.