The Ethical Gravity Thesis: Marrian Levels and the Persistence of Bias in Automated Decision-making Systems

Atoosa Kasirzadeh
University of Toronto
Australian National University
atoosa.kasirzadeh@anu.edu.au

Colin Klein
Australian National University
colin.klein@anu.edu.au

ABSTRACT

Computers are used to make decisions in an increasing number of domains. There is widespread agreement that some of these uses are ethically problematic. Far less clear is where ethical problems arise, and what might be done about them. This paper expands and defends the *Ethical Gravity Thesis*: ethical problems that arise at higher levels of analysis of an automated decision-making system are inherited by lower levels of analysis. Particular instantiations of systems can add new problems, but not ameliorate more general ones. We defend this thesis by adapting Marr's famous 1982 framework for understanding information-processing systems. We show how this framework allows one to situate ethical problems at the appropriate level of abstraction, which in turn can be used to target appropriate interventions.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics** → **Governmental regulations**; • **Computer systems organization** → **Embedded systems**.

KEYWORDS

Ethics of Artificial Intelligence, Politics of Artificial Intelligence, Ethical Artificial Intelligence, Ethical Machine Learning, Algorithmic Bias, Algorithmic Fairness, Justice, Philosophy of Artificial Intelligence

ACM Reference Format:
Atoosa Kasirzadeh and Colin Klein. 2021. The Ethical Gravity Thesis: Marrian Levels and the Persistence of Bias in Automated Decision-making Systems. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3461702.3462606

1 INTRODUCTION

1.1 What would it mean to make ethical AI?

You are a programmer. Your job is to make a system that determines eligibility for high-interest payday loans. Your employer is predatory. The customer base is vulnerable. To maximize profit, you must identify individuals who will never quite be able to repay their loan—guaranteeing a steady stream of income for your employer—but for whom that burden will not result in bankruptcy, suicide, or other drastic ways that the desperate find to discharge their debt.[1] Having been deeply enmeshed in the literature on ethical algorithms, you decide that you ought to make the system more ethical. How would you do that?

The ambiguity in that question should be obvious. On the one hand, you can make sure that you do not introduce any new ethical problems. You might avail yourself of any number of proposed metrics for measuring algorithmic fairness [6, 14, 20, 26, 28, 50], and ensure you optimize one. You might blind your dataset to avoid information about protected attributes. You might ensure that you provide a detailed 'model card' specifying all of the information used in building your model [37]. On the other hand, what you make will lead to substantially immoral outcomes. That is true no matter how careful you are: the design goals of the system place a fundamental upper bound on how ethical the result could be.

There is now a considerable literature on the negative ethical consequences of these automated decision-making systems.[2] A good deal of this literature also contains specific technical recommendations. Yet while many ethical problems have a technological source, it is not at all clear that they have technological solutions.

More generally, discussions of automated decision-making systems often contain a wealth of different concerns, with the relationships between them unclear. Some authors are concerned with the ethical deployment of technology. Others are concerned with details of algorithms or other mathematical objects. Still others care about the details of particular implementations: where datasets come from, or who gets paid for annotations. The relationship between these different projects is often obscure, and hence the effectiveness of particular interventions is difficult to evaluate.

We will argue for a particular thesis about the relationship between these different ethical projects, what we call the *Ethical Gravity Thesis* (EGT). Making EGT precise will take a bit of doing:

AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462606

[1] Mayer [32] reviews the ethical case for thinking of payday loans as exploitative across a variety of theories of exploitation. See Melzer [34] for a good review of the overall cashflow problems that are worsened by payday lending. If you sympathize with the neo-liberal view that properly-regulated payday lending provides a valuable source of liquidity, stipulate that this is one of the problematic ones that gives payday lending a bad name.

[2] For useful recent reviews, see Coeckelbergh [8], Kearns and Roth [24], Mittelstadt et al. [38], Torresen [48], Whittlestone et al. [52]. We use the broad term 'automated decision system' to cover any use of computers to make decisions in ethically relevant contexts. This is meant to be neutral and inclusive, and so include terms like 'Artificial Intelligence', 'Machine Intelligence', 'Machine Learning', and 'Algorithmic decision-making', as well as more specific terms of art that refer to particular methods for implementing decision-making. Different terms carry different connotations; ours is intended to be as neutral as possible, and in particular not to interfere with the sense of 'algorithmic' detailed below.

for now, take it as the thesis that ethical problems at higher levels of analysis of an Artificial Intelligence (AI) system cannot be reliably ameliorated by interventions at a lower level of analysis. The toy example with which we began is a stark illustration of that principle. What follows will give a more nuanced model. Before we spell out the detailed goals, however, some background is in order.

1.2 Background

Automated decision-making has impacted various aspects of public and private life almost since the advent of digital computers. These systems are now prominent in a variety of ethically important domains. Judges use algorithms to decide whether defendants awaiting trial should be detained or released [7, 39]. Health care providers use commercial AI to guide healthcare decisions [3, 43]. Law enforcement uses facial recognition algorithms embedded in public spaces to facilitate continuous surveillance [16, 17, 25].

These uses all come with both ethical costs and benefits. Yet there are also troubling signs of ethical problems that arise and are specific to automated decision-making systems.

There is growing global concern, in some cases empirically confirmed, about the harmful and disruptive impacts of automated decision systems. Research has shown that some automated job search advertisements for highly paid positions are more likely to be presented to men [29], risk scores in recidivism prediction are interpreted to be biased against Blacks and some other disadvantaged groups [2, 15], facial recognition systems perform less accurately on recognizing faces of women and Black individuals [5], and algorithms embedded in health care decision-making show systematic discrimination against Blacks [40].

We ought to care about the ethical considerations of the use of automated decision systems, and we must anticipate and mitigate the problematic and unjust effects stemming from their deployment. Despite that general agreement, different discussants and institutions often formulate such ethical problems at distinct levels of abstraction, and consequently offer solutions that are not obviously commensurate.

On the one hand, there are already several dozen national and international ethical guidelines for the development, deployment, and governance of AI algorithms using high-level, non-technical language [23]. These guidelines are often vague and lack detailed mechanisms for reinforcing ethical principles [19]. On the other hand, there are technical proposals for how to make algorithms 'ethical' [24]; for example, there is a heated debate and a substantial technical literature about how to make algorithms transparent and fair [1, 6, 26]. Yet focus on technical solutions can obscure the role of social and political change in certain realms. It can also invite so-called 'ethics washing,' where companies make minor technological changes to preserve their public image while continuing business as usual [4, 51].

It is difficult to make automated decision-making systems ethical without knowing what that task amounts to. The literature urgently needs a framework for the systematic treatment and organization of ethical concerns, including the dependency relationships between them. The present paper is intended to make steps in that direction.

1.3 The Claims and the Plan

Our contribution to the ongoing debate will be twofold. First, we begin by adapting Marr's framework for understanding complex information-processing systems to discuss different levels of analysis for automated decision-making [31]. Marr's framework has been extremely influential in cognitive science, but remains underutilized in work on automated decision systems. The framework is useful in part because it implies asymmetric dependence relationships between levels: in general, higher levels of analysis provide strong constraints on lower ones.

Second, we use our framework to claim evidence for the Ethical Gravity Thesis. The EGT claims that ethical considerations that appear at a higher Marrian level of analysis cannot be overcome, in any robust way, by interventions at lower levels. Combined with a Marrian analysis, the EGT thus provides a groundwork for a systematic analysis of automated decision-making. In many cases, this shows that what appear to be distinct problems are merely symptoms of more basic ethical worries. It also opens the door to broader considerations about institutional design and algorithmic fairness, and the ways in which the two interact.

2 LEVELS OF ANALYSIS

The Ethical Gravity Thesis makes a claim about dependence between different levels of analysis of an automated decision system. What are those levels? To answer, we start by giving a quick overview of Marr's influential analysis of complex information-processing systems [31]. Marr identified several levels that are common to the analysis of any computational system. While his framework was originally developed to account for different levels of analysis in cognitive neuroscience, it has had a broad influence across many fields concerned with computational explanation.

Marr posits four levels of analysis at which any computational process can be approached. We discuss each, drawing the parallels between his target and our own.

Two caveats are in order (and will be fleshed out as we go). First, automated decision-making systems are made by humans, while Marr was concerned with systems shaped by natural selection. Part of the power of Marr's framework is that it is applicable to both natural and artificial systems. In either case, we take it that the levels represent a series of distinct ways of *analyzing* a system, rather than a guide to the temporal order of discovery or of creation. Second, the levels of analysis are not meant to be the levels of *blame or responsibility*. Higher levels tend to be analyzed in terms of public or corporate policy, while lower levels in terms of the actions of individual programmers. But the fact that a moral problem is *introduced* by a particular level of analysis does not show that any particular set of people is responsible for that problem. Blame often depends on the compounding effects of many distinct factors (indeed, as we will discuss later, it is possible for ethical problems to arise completely blamelessly).

2.1 The Functional Level

Marr's levels were initially offered as a way to understand and explain early visual processing. The highest level of analysis of this—or any other complex phenomenon—is the characterization of

what goals the system aims to achieve. We call this the functional level of analysis.[3]

We focus (as is traditional) on Marr's discussion of edge detection, one of the robust visual phenomena that early vision must accomplish.[4] For Marr and for his study of vision, these are drawn from the various phenomena of early vision, revealed by ordinary experience and simple psychophysics. Marr begins his analysis by noting phenomena that we are all familiar with—the appearance of subjective contours, the dominance of pattern and surface in the perception of shapes, or the ability of lines to 'pop out' in visual experience given all manner of subtle cues [31, Ch 2.1]. These phenomena are given in experience.

The functional level of analysis of automated decision-making systems identifies the overall goals that the system is meant to accomplish. This, as it were, a pre-technical description—the sort that could be offered to executives, marketing, or sales. This might be 'balance the risk of recidivism with the need to parole', or 'figure out who to give loans to in order to maximize profit.'

The functional level of analysis specifies what a system must do to count as successful. Of course, though we might sketch the function of a system in a short elevator pitch, the full functional specification of a system depends quite a bit on context. In the case of edge detection, not any way of detecting visual edges will do: edge-detection is in the service of survival, and the world places very strict limits on what will work. Similarly, nobody gives out loans in a vaccuum: the practice of loan-giving is made possible by a host of regulations, expectations, and social practices that shape and contour it in various ways. We include all of these contextual effects in the functional level of analysis.

2.2 The Computational Level

Return to Marr's explanation of early vision. Having decided to explain edge detection, we make further progress by explaining the particular computational function that would result in edge detection. That is, given retinal input, what is the mathematical characterization of the function that goes from that input to give us edges in the scene? Marr notes that edge detection can be characterized as the discovery of zero-crossings in the second derivative of the two-dimensional array of intensities provided by the retinal image [31, 54]. This function can be calculated by a simple combination of mathematical functions that work as filters to pick out the zero-crossings.[5]

Note that there are many (perhaps indefinitely many) mathematical functions that could be said to perform edge detection. Marr's particular choice of filters is dictated in part by his analyses at the functional level above. We perform edge detection robustly across different intensities and spatial scales. That is one of the phenomenon that the computational story must capture.

Conversely, the computational formulation is still quite abstract. The function identified could be implemented just as well in silicon as it could be in neural tissue. Again, Marr was influential in part because his theory made clear the abstract links between cognitive neuroscience and computer science, providing an important theoretical foundation for work on artificial intelligence.

Any automated decision-making project can also be understood as attempting to perform a certain mathematical task. Determining which loans to make might require the simultaneous optimization of one or more equations. Recognizing objects in pictures is formally equivalent to untangling manifolds in a very high-dimensional space [12, 13].

Some computational processes are learning processes: the real task is finding an appropriate mathematical function from data to another function that will provide answers in production. In each case, however, we can view the functionally defined problem through a mathematical lens—and, indeed, we must do so if we are to program a computer to find the answer we seek.

2.3 The Algorithmic Level

Mathematical functions are timeless. To get a computer to *do* something, we need an algorithm: that is, a "set of rules or directions for getting a specific output from a specific input" [27]. Algorithms proscribe a sequence of steps and operations that must use time and other resources.

The algorithmic level of analysis is constrained by empirical facts about those resources. One might detect lines (for example) by building a big set of specific feature detectors, one for each possible combination of intensities that could fall on the retina. That would solve the mathematical problem neatly, but would require vast amounts of storage capacity, far more than is plausible. Algorithms that iterate over the visual field in a serial fashion would similarly take too long to be biologically useful. So instead, according to Marr, early vision evolved to use various tricks that also proved useful to later computer vision researchers: local pooling of information, compression of information via the use of adaptive local filtering, and so on.

Automated decision-making also faces algorithmic choices. The specific choice of algorithm (a Greedy algorithm or Dijkstra) to optimize a given mathematical function can make a drastic difference to how accurately or quickly the function is computed. It has long been known, for example, that even poor-quality linear regressions can often outperform humans on simple decision tasks at trivial computational cost [11], whereas further improvements in accuracy require rapidly increasing investment of resources.

An important feature of the algorithmic level, as Marr was well aware, is the way in which data is represented. The efficiency of available algorithms depends on the primitive representations upon which algorithms operate [53]. To take a simple example: if I represent my loan applicants' data as points in a high-dimensional continuous feature space, I will have available many more, and more accurate, classification algorithms than I would if I represented each applicant as a simple dictionary of categorical properties.

[3] Many discussions of Marr focus on only three levels; what we call the 'functional' level is often elided because it is taken as the starting point of cognitive neuroscience. We think it is clear that a superordinate functional level exists–see especially the top of figure 6.1 in [31]. The precise content of Marr's framework presents interpretive challenges [44]. As with many who write on Marr we claim inspiration, not tight textual fidelity.

[4] The details of Marr's theory of edge detection are historically important, but many of the details have been superseded. We follow tradition in presenting Marr's theory, but for an up-to-date discussion of the functions and constraints of early vision in mammals, see e.g. [47, Ch 12].

[5] In Marr's famous formulation, we can discover these zero-crossings by using the filter $\nabla^2 G$, where ∇^2 is the Laplacean operator and G is a two-dimensional Gaussian filter.

2.4 The Implementational Level

The first three levels were all hardware-agnostic. Though Marr's goal was to analyze how edge detection is done by early visual cortex, the computational and algorithmic levels did not require knowing anything at all about neurons or how they work. Even the time and space constraints on the algorithmic level can simply be taken as primitive facts for the purposes of algorithmic analysis.

For Marr, the implementational level shows how the algorithmic level is embodied in particular hardware. It shows, for example, how the center-surround architecture created by patterns of excitatory and inhibitory cells can implement the primitive filters identified at the previous steps. Whereas any of the preceding levels could have been implemented in silicon, the implementational level for Marr is firmly tied to biological detail.

For example, there are many algorithms that can learn to recognize faces and the emotional states they represent (for a survey see [33]). But to run them for predicting the emotional states of unobserved faces, one needs to train the algorithms on a dataset of faces with associated labels representing their emotional states. How many faces does that dataset contain? Where did it come from? Are the exemplars appropriately distributed over different genders? Different races? Different ages? What is the basis of the emotional groupings? How may different views does each face have? Are they contrast-balanced? How reliable is the labeling? And so on and on. By and large, algorithms work with whatever they are given; the details of the implementation thus determine how *well* the algorithm serves its purpose.[6]

In addition to the implementational details of the task at hand, one might also include here all of the details that are necessary for (e.g.) compliance and auditing of the program—details like the collection of logs, the ability to vary parameters while in production, and so on. These are features of particular uses of algorithms, and so features of implementation.

2.5 The Overall Picture

Figure 1 collects up the discussion so far. The boundaries between the levels are sometimes fuzzy, but in general, we submit, the schema holds surprisingly well across a variety of analytic tasks.

The center column of figure 1 contains additional structure in the form of arrows between levels. We have suggested that the levels do not vary freely; there are relationships of constraint among them. It is to those arrows that we now turn.

[6]Some readers of Marr might balk at including details like datasets at the implementational level. There is a traditional reading of this level of Marr on which it is concerned solely with bottom-level hardware—neurons for brains, silicon for automated decision-making. We resist this reading for several reasons. First, we think it is rare that the pure hardware level represents an interesting level of ethical analysis. Second, there is an important distinction between algorithms considered in the abstract and the particularized instances of programs and data upon which algorithms are run. Marr's analysis (we claim) clearly applies to the abstract sense of algorithms, which leaves particular details to the lower level. Third, the natural analog of datasets in the early visual case is facts about incoming light and its interaction with receptor physiology. Fourth, discussions of the implementational level, both in Marr and elsewhere, are not mere catalogs of squishy bits but themselves conducted at a level of abstraction that brings out their crucial functional characteristics. Hence, we claim, what is important about the implementation is *anything* that is relevant to the instantiation of an algorithm as a token process, not just the hardware.

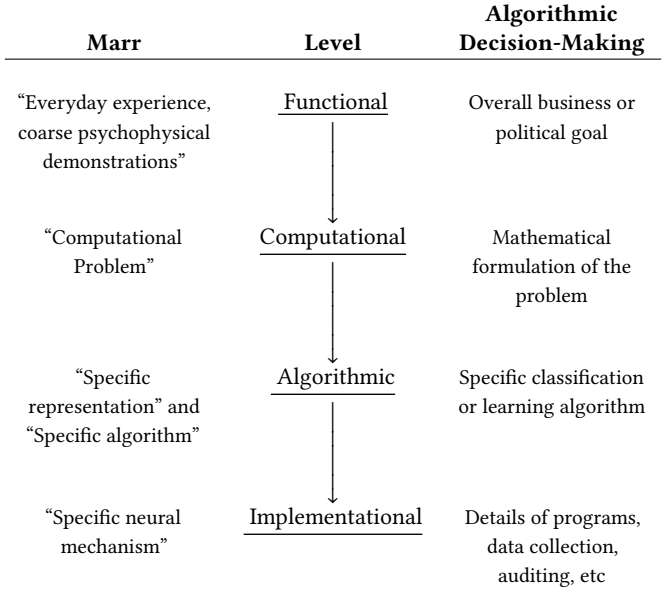

Figure 1: The relationship between Marrian levels of analysis for complex information-processing systems.

3 TWO ARGUMENTS FOR EGT

The Ethical Gravity Thesis can now be stated more precisely. If an ethical problem arises at one level in the Marrian hierarchy, one should expect it to persist at all lower levels. Lower levels can introduce new ethical problems, but one should not expect them to get rid of problems inherited from higher levels. This is true even when people responsible for creating or maintaining the lower levels have a sincere and strong incentive to ameliorate ethical problems.

Put concretely: insofar as making a payday loan system is ethically problematic at all, the ethically problematic features do not go away when you make a particular version of that system. The lower levels might tack on additional ethical problems, but cannot avoid the ones inherited from the functional level.

A few clarifications are necessary. First, The EGT is phrased in terms of reasonable expectations. There may be isolated instances in which lower levels solve ethical problems introduced at higher ones, either intentionally or accidentally. EGT claims that such instances are neither frequent nor particularly robust. Second, when we talk about 'ethical problems', those should be understood extensionally. That is, a decision-making problem is ethically problematic when it has ethically problematic *effects*—people are harmed or exploited, inequality is exacerbated, and so on. We follow most of the current literature in assuming that bad *intentions*, either by individuals or by groups, are not necessary for a process to be ethically problematic. Bad intentions may play an important role in making bad effects robust and stable (about which more shortly). But the EGT is concerned with bad consequences.

In support of EGT, we offer two arguments: the *realization argument* and the *institutional argument*. The realization argument claims that realization relationships between levels—the arrows in figure 1—require lower levels to faithfully perform the task as

specified at higher levels. This means that an ethical problem that can be identified at one level should be expected to be preserved by all lower levels. The institutional argument claims that active social and political forces make even *un*successful realizations—that is, realizations that only fulfill some of the conditions set by higher levels—relatively infrequent and difficult to maintain.

The overall form of the argument is thus a disjunctive syllogism. Either lower levels successfully realize higher ones or they do not. If they do, then ethical problems are preserved. If they do not, and their failure to completely realize a higher level manages to solve some ethical concern, then there will be strong pressures to correct that failure. Either way, we should expect that, in the long run, ethical problems at higher levels will be preserved at the lower levels. Together this establishes the EGT.

3.1 The Realization Argument

The arrows in figure 1 represent relationships of realization.[7] Realization is a relationship of asymmetric necessitation: B realizes A just in case being B is a way of satisfying A but A does not necessarily require B.

To make the notion of realization relation more concrete, consider a simple example. I have a list of books that I need to alphabetize. That is a functional-level statement of the problem. To solve this, I need a function which takes a list of book titles and returns a lexically sorted list. Doing this is *a way of* satisfying the functional-level statement of the problem. The functional-level problem does not determine the computational-level function, however, even for such a simple case. Sorting functions might differ on how they deal with duplicates, or non-Latin characters. Hence there is a relationship of asymmetric necessitation: performing a particular sort function is a way of alphabetizing books, but the requirement to sort does not necessitate any particular function. Similarly so down the chain: having specified my function, I can use many different algorithms (Selectionsort or Quicksort or…) to perform the task. And having chosen an algorithm, I still have flexibility on the programming language and hardware that I use to implement it. (Indeed, I may decide to implement a useful algorithm using pen and paper rather than a computer!)

Given this, the realization argument for EGT is straightforward. Suppose an ethical problem with negative impacts appears at a higher level of analysis. Then the solution to this problem at a lower level of analysis is a way of satisfying the ethical demands appearing at the higher level. The next lower level of analysis must asymmetrically necessitate the higher level: that is just what it is to realize the higher level. Part of asymmetric necessitation is preserving extensional fidelity to the activities at a higher level. But then the lower level must just have the same effects as the higher level, just in a more specific manner.[8] By induction, the problem will persist not just at the next lowest level but at all lower levels.

Moreover, as the realization relationship is one of asymmetric dependence, there is considerable flexibility in how any level is realized at a lower level. This means that lower levels can still *introduce* new ethical problems. A perfectly reasonable facial recognition algorithm implementing an ethically anodyne task might still go astray if trained on a biased dataset. So lower levels preserve the ethical problems, but can add more besides, which is just what the EGT says.

3.2 The Institutional Argument

The realization argument establishes that ethical problems percolate downward when realization is satisfied. Computers make mistakes. Say that a higher level is *partially realized* by a lower level when the activity of a lower level necessitates the prescribed effects of the higher level across some but not all contexts. Our sorting algorithm, for example, might realize our sorting function except when Cyrillic characters appear (in which case it crashes).

Those who work on the development of automated systems know how common partial realization is. It is particularly common in complex systems, and morally-laden decision-making systems are likely to be especially complex. Hence (one might hope) negative effects that would otherwise occur due to constraints at higher levels might be partially ameliorated by lower levels. This might occur by accident. It might occur via the sneaky actions of a clever and motivated programmer who is determined to undermine the system. Or it might occur by sheer diffidence and inertia at the implementational level. However it happens, one might hope that these 'mistakes' could fix some of the otherwise problematic aspects of a system.

This is a tempting thought, but we think that it is unlikely to be particularly effective in practice. There are strong *institutional* constraints on the creation and maintenance of automated decision-making systems. An institutional constraint is any set of procedures or processes that tends to move partial realization to full realization by (e.g.) correcting errors, weeding out inefficiencies, and so on.

The key point here is that partial realization is a kind of *failure*: it occurs only when the putative way of doing some activity does not completely align with that activity across all relevant contexts. There are a variety of institutional constraints that exist in order to catch errors of this sort and reverse them. Institutional constraints thus provide a natural limit to the effectiveness of interventions at a lower level.

We consider three sorts of institutional constraints, though these are meant to be illustrative. The broader point is that the existence of institutional constraints means that substantial partial realization—and especially ethically relevant partial realization—should be relatively rare and fragile.

First, automated decision systems are complex software projects. Modern software projects, whether commercial, governmental, or open source, are usually embedded within a fairly complex system of procedures and practices that are designed to catch and mitigate

[7] A note for aficionados: we take the realization relationship to be a 'flat' rather than a 'leveled' one [18]. That is, the distinct levels of analysis pick out different properties of the same system, rather than lower levels picking out functional or spatiotemporal parts of higher-level systems, as they do in mechanistic explanations [9]. Mechanistic explanation is likely to play a further, important role in explanations of implementation, but higher levels do not have interesting spatial structure, and the computational level arguably does not even have temporal structure. For a flat reading of Marr in the computational domain, see Piccinini [42, 98ff]. See also footnote 8.

[8] In other words, we view realization as a kind of determinable-determinate relationship, along the lines of Yablo [54]. Hence lower levels must do everything that higher levels do, and more besides (because it is done in a particular manner).

bugs. This includes practices such as widespread unit tests, code reviews, bug tracking and version control. These practices are meant to minimize the occurrence of partial realization at the algorithmic and implementational levels. They are not uniformly successful (obviously so). Yet their primary purpose is to catch and eliminate instances of partial realization. Ethically relevant partial realization is just one of many sorts of bugs that it might weed out.

Second, automated decision systems are complex social and economic entities. Most of the people who would be in a position to effect partial realization also have substantial incentives to avoid it. There is a rich philosophical tradition on the concept of *complicity* that analyses the moral obligations of individuals in ethically problematic organizations. This literature also acknowledges, and lives alongside, a rich social-psychological tradition demonstrating how difficult it is for individuals to actually make a meaningful difference in these situations, especially under conditions of diffuse responsibility.[9]

In addition to social constraints (broadly conceived), there are often substantial economic incentives to avoid partial realization. Programmers who do not do their jobs can be fired; firms who do not deliver lose contracts.[10] Even when these incentives do not prevent subtle resistance by individuals, they may give *other* individuals reasons to detect and undo attempts at partial realization.

Third, automated decision systems, as socio-technical entities, themselves arise from, are embedded in, and are maintained by social and political forces. To return to our initial toy case, the existence of high-interest payday loans does not occur in a vacuum. Nor, one supposes, are high-interest payday loan businesses generally run by people with a strong concern for economic justice. So the problem is not accidental: many people involved in the system do not want it to be better.

We phrase these processes in an intentional idiom, but in practice many discriminatory effects can occur through what are known as structural biases [21]. Structural biases are embodied in norms and institutional designs, and do not require any individual to have objectionable attitudes. They provide a kind of inertia that ensures that institutions and individuals do not, and often cannot, deviate from that tacit norm. For example, Crawford and Paglen [10] provide a a realistic picture of how software arises in a social context. Their focus is on image recognition and the role of apparently neutral datasets in perpetuating certain kinds of disadvantage. Their analysis shows that even apparently innocuous uses of image recognition are embedded in a vast social and political landscape, one that effectively works behind the scenes to stabilize seemingly arbitrary features of decision-making systems.

These different institutional pressures may have differential effects at different levels of analysis. Broadly speaking, technical and procedural institutions work to correct partial realization at the implementational and algorithmic levels, while economic and social constraints are more operative at the functional and computational level. Such institutions do not need to be entirely effective; instead, they need only to be effective enough that we should not expect partial realization (whether intentional or accidental) to be particularly robust.

The proof is thus complete. Full realization ensures that ethical problems are inherited by lower levels. Partial realization does not last for long. Hence ethical problems are either faithfully preserved by the lower level, or else the fix is unlikely to last long. That is just what is claimed by the EGT.

4 ETHICAL GRAVITY: CASES AND CONSEQUENCES

4.1 Recidivism Prediction

To illustrate the EGT further, we conclude with several examples.

Few automated decision-making systems have been the focus of as intense a discussion as COMPAS. The recidivism risk-prediction algorithm at the heart of COMPAS that was the target of a now-famous 2016 ProPublica analysis of its use in Broward County, Florida. That analysis showed that COMPAS embodied certain kinds of racial bias [2]. Although race was not an explicit feature in the algorithm, it suggested that Black prisoners are substantially more likely to be classified as high risk. Moreover, Angwin et al. [2] showed that among defendants who ultimately did not re-offend, Blacks were more than twice as likely as whites to be labeled as risky by the algorithm.

COMPAS shows several distinct ways in which problems perpetuate downward. First, as subsequent discussion has shown, there is a difficulty that arises at the computational level. Assuming that fairness is a desideratum, and that minimizing recidivism is another desideratum, then there exist distinct tradeoffs between different fairness metrics [6, 26]. This is a mathematical fact, but it obtains because of broader societal facts: various intuitively plausible fairness metrics can only be jointly satisfied if populations of offenders are balanced in various ways that they are not, in fact, actually balanced. The reasons for this imbalance are broader functional-level injustices.

As such, problems at the functional level assure that intuitively plausible desiderata at the computational level cannot be jointly satisfied. So there is a tradeoff at the computational level which, in turn, perpetuates further ethical problems down the line. Given that, no algorithm or no implementation could hope to ameliorate the ethical issues that are ultimately inherited from the functional level: the constraints at the computational level forbid an effective solution.

Much of the discussion around COMPAS has focused on these purely formal considerations. However, we think that this is a case where institutional facts are also notable, and provide a good illustration of the EGT. The subsequent debate and argument over COMPAS makes it arguably the most-studied case in discussions of algorithmic ethics. What is less often mentioned is that Northpoint's (now a marquee of Equivant) COMPAS software is still for sale, and still appears to be widely used. Their FAQ asks "Is the COMPAS algorithm racially biased?" and responds "No, COMPAS

[9] For an excellent philosophical review of complicity, see Lepora and Goodin [30]. For obedience, the *locus classicus* is of course Milgram [35], though his [1974] considers a broader set of parameters. Staub [46] gives a good modern review of Milgram and related work on bystander effects.

[10] "Before the war the pacifists had more than once explained to us that a country that has been invaded must refuse to fight and engage in passive resistance. That's easy to say: but in order for this resistance to be effective the railroad workers would have had to refuse to let the trains run and the farmers to work the fields. The victor would have been inconvenienced but he could have supplied himself from his own country; however, the occupied country would certainly have perished in short order." [45, 11]

is designed to assess numerous factors, but race is not even considered when a COMPAS score is developed."[11] That this is the entirety of the answer says something important about Equivant's *customers*: whoever is in charge of procurement considers this to be a perfectly adequate response.

This should not be a surprise. Equivant's customers are criminal justice agencies. Criminal justice agencies in the US do not, on the whole, have a reputation for progressive attitudes towards racial inequality. Rarer still is the criminal justice agency who would rank undoing racial inequality over (say) avoiding the social and political consequences of a serious crime committed by a paroled offender. COMPAS continues to be popular, then, precisely because it successfully embodies the values of the institutions that use it. But these institutions are precisely the institutions that perpetuate and partly constitute the injustices at the functional level.

Hence the problem with COMPAS is not merely formal, though the formal features prevent an easy fix. Instead, considerable political will at the functional level ensures that the status quo is perceived as perfectly acceptable, at least to those with the power to select and implement risk-management algorithms; conversely, what might seem like lower-level 'fixes' will be seen as *flaws*, and be selected against accordingly. That is, recall, just what the institutional argument for EGT claimed.

4.2 Allocation of Scarce Health Resources

Some algorithms reflect ethical problems at a broad societal level. Yet the EGT also allows that lower levels might create *new* ethical problems, either directly or unwittingly.

Consider Obermeyer et al. [40]'s recent analysis of healthcare allocation systems. These automated systems attempt to allocate scarce and costly healthcare resources to those who would benefit from them the most. This broad goal corresponds to the overall goal at the functional level as illustrated in figure 1. This is not (let's assume) an ethically problematic objective. Obermeyer et. al. showed that the algorithm in charge of this resource allocation task was largely driven by translating the problem in terms of predictions of healthcare costs at the computational level. At first glance, this is unsurprising: as they note, costs and health needs are highly correlated.

Yet in the US there is also a background correlation between race and healthcare costs: on average, Blacks receive less care for a given level of need than do Whites. As such, the algorithm tends to dramatically underestimate the need for additional care among Black patients. Together, these facts mean that Black patients were thus less likely than Whites to be referred to support programs for patients with complex medical needs. Only 18% of patients that the algorithm assigned to receive extra care were Black; the figure for an algorithm would be closer to 47% if it were unbiased. Back to the EGT argument, this study shows that the ethical concerns for the allocation of health care resources persist at the lower levels of automated decision-making. Yet even this decision-making system can instantiate the institutional argument as follows.

[11] From the Equivant FAQ, https://www.equivant.com/faq/, accessed 14 August 2020. For claims of wide use, see same FAQ, "Also, we have over 100 supervision, inmate classification, and risk/needs assessment systems that include Department of Corrections in seven states along with numerous implementations in pretrial, probation, and sheriff/jail offices across the U.S."

The problem comes from using what seems like a plausible proxy for healthcare costs that interacts poorly with other background facts. As they put it, "accurate prediction of costs necessarily means being racially biased on health." [40, 450]. The discussion around this phenomenon, so far as we know (and unlike in the case of COMPAS), assumes that this bias is straightforwardly bad. There is a subset of patients for whom the algorithm performs poorly. That poor performance is bad for the individual patients and bad from an economic point of view: untreated medical conditions tend to get worse, and more costly.

The ethical problems thus arise not at the functional level but at the computational and algorithmic levels. They are a specific instance of what Passi and Barocas [41] call the issue of problem formulation. Building an automated decision-making system involves a series of what they call "difficult translations." These include the move from policy objectives to specific algorithms, the choice of objective functions, and the specific, quantitative features used for problem prediction. One might worry that certain kinds of ethical properties cannot, in principle, be satisfactorily mathematized [49]. But one might also worry that, even among the properties that can be satisfactorily mathematized, there are still additional ethically laden questions that must be answered.

4.3 Facial Recognition

Facial recognition technology has been in use since the 1960s [22], but technological advancements have led to widespread deployment. Particularly concerning from an ethical point of view, is its use by law enforcement. As in the case of COMPAS, there are obviously biased uses of facial recognition that can arise at the functional level. However, facial recognition is also used for a variety of apparently innocuous tasks like tagging pictures on social media or unlocking smartphones.

These more innocuous uses might appear to avoid novel ethical issues. However, many of the datasets used to train otherwise innocuous technology themselves have differential representation of classes, which can lead to poor accuracy on members of racial minorities [25]. This is obviously problematic when used for criminal justice purposes [22]. However, such disparities can introduce novel ethical issues even in otherwise straightforward contexts.

Consider, for example, the finding that some gender prediction software has the lowest accuracy rate for Black women [5]. This differential accuracy might be seen as a form of micro-aggression, or simply lead to lack of uptake of a valuable new technology by members of a vulnerable group.

Much of the discussion around dataset biases in facial recognition runs together these possibilities. We think the EGT is useful because it allows one to partition out different ways in which dataset biases might matter. They might matter because they are simply part of the implementation of a fundamentally unjust or biased system. In that case, worry about particular datasets is important because they show how ethical problems are *exacerbated* by implementations. However, the ethical problems arising at higher Marrian levels cannot, on their own, be fixed merely by fixing the dataset.

On the other hand, dataset biases might introduce fundamentally new ethical problems in an otherwise neutral decision-making

system. In this case, it is worth direct intervention on the dataset itself.

5 CONCLUDING THOUGHTS

In this paper, we have demonstrated how to apply Marr's framework for understanding information-processing systems to automated decision-making systems. Using this framework, we have also argued for the Ethical Gravity Thesis: that ethical issues that arise at higher levels of the Marrian hierarchy will necessarily be inherited by lower levels. To put it in the starkest terms, some automated decision-making systems thus cannot be made ethical, even in principle. One can at best minimize the damage that they cause by not introducing new problems in their implementation.

The EGT is humbling. Yet it is not meant to be pessimistic. We conclude on two (comparatively) more optimistic points. First, as it is possible to introduce ethical dilemmas at lower levels, there are important design considerations that need to be explored at each step down the chain. Indeed, one of our goals in writing this paper was to draw the distinction between ethical problems that arise in the course of implementation from ethical problems that arise because an automated system is embedded in, and in service to, a fundamentally unjust social and political environment. The latter cannot be solved by technical acumen; such failures should not lead to pessimism about addressing the former.

Second, we have mostly focused on the realization relationship, which moves downward in the Marrian hierarchy and which is the source of the EGT. There is arguably a converse relationship of *constraint* that flows upward. Various choices made at higher levels are made against a background of implicit or explicit resource constraints introduced at lower levels. Some of these constraints are insuperable: in terms of computational complexity, for example, if the most ethical algorithm is NP-hard, we cannot build a practically useful system.

Yet many constraints are economic, or only contingently technical. For example, if our financial conditions permit us to only use freely accessible, publicly-available datasets, and these datasets are all biased, then the solution is to find more financial budget or to encourage the development of better freely accessible, publicly-available datasets. Removing constraints at lower levels cannot solve ethical problems directly: that is what the EGT shows. But removing constraints can solve ethical problems *indirectly*, at least in some cases, by removing constraints on higher levels that give rise to ethical problems down the line. Examining the dual relationship between realization and constraint may thus show an additional pathway for ameliorating ethical problems raised by automated decision-making.

6 ACKNOWLEDGEMENTS

This project was supported by the Humanising Machine Intelligence Grand Challenge at the Australian National University.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Julia Angwin, Larson Jeff, Mattu Surya, and Kirchner Lauren. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals and It's Biased Against Blacks. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] Andrew L Beam and Isaac S Kohane. 2018. Big data and machine learning in health care. *Jama* 319, 13 (2018), 1317–1318.

[4] Elettra Bietti. 2020. From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 210–219.

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[6] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[7] Angèle Christin, Alex Rosenblat, and Danah Boyd. 2015. Courts and predictive algorithms. *Data & Civil Right: Criminal Justice and Civil Rights Primer* (2015).

[8] Mark Coeckelbergh. 2020. *AI Ethics*. MIT Press, Cambridge.

[9] C.F. Craver. 2007. *Explaining the Brain*. Oxford University Press, New York.

[10] Kate Crawford and Trevor Paglen. 2019. Excavating AI: The politics of images in machine learning training sets. *Excavating AI* (2019). https://www.excavating.ai.

[11] Robyn M Dawes. 1979. The robust beauty of improper linear models in decision making. *American Psychologist* 34, 7 (1979), 571–582.

[12] James J DiCarlo and David D Cox. 2007. Untangling invariant object recognition. *Trends in Cognitive Sciences* 11, 8 (2007), 333–341.

[13] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. 2012. How does the brain solve visual object recognition? *Neuron* 73, 3 (2012), 415–434.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[15] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, New York.

[16] Andrew Guthrie Ferguson. 2019. *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York University Press, New York.

[17] Kelly A Gates. 2011. *Our biometric future: Facial recognition technology and the culture of surveillance*. New York University Press, New York.

[18] Carl Gillett. 2003. The metaphysics of realization, multiple realizability, and the special sciences. *The Journal of Philosophy* 100, 11 (2003), 591–603.

[19] Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines* 30 (2020), 99–120.

[20] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems* 29 (2016), 3315–3323.

[21] Sally Haslanger. 2015. Distinguished lecture: Social structure, narrative and explanation. *Canadian Journal of Philosophy* 45, 1 (2015), 1–15.

[22] Anil K Jain, Brendan Klare, and Unsang Park. 2011. Face recognition: Some challenges in forensics. In *Face and Gesture*. 726–733.

[23] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[24] Michael Kearns and Aaron Roth. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, New York.

[25] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801.

[26] Jon Kleinberg. 2018. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. 40–40.

[27] Donald E Knuth. 1977. Algorithms. *Scientific American* 236, 4 (1977), 63–81.

[28] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4069–4079.

[29] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science* 65, 7 (2019), 2966–2981.

[30] Chiara Lepora and Robert E Goodin. 2013. *On complicity and compromise*. Oxford University Press, Oxford.

[31] David Marr. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge.

[32] Robert Mayer. 2003. Payday loans and exploitation. *Public Affairs Quarterly* 17, 3 (2003), 197–217.

[33] Dhwani Mehta, Mohammad Faridul Haque Siddiqui, and Ahmad Y Javaid. 2018. Facial emotion recognition: A survey and real-world user experiences in mixed reality. *Sensors* 18, 2 (2018), 416.

[34] Brian T Melzer. 2011. The real costs of credit access: Evidence from the payday lending market. *The Quarterly Journal of Economics* 126, 1 (2011), 517–555.

[35] Stanley Milgram. 1963. Behavioral Study Of Obedience. *Journal of Abnormal and Social Psychology* 67 (Oct 1963), 371–378.

[36] Stanley Milgram. 1974. *Obedience to Authority: An Experimental View*. Tavistock Publications., London.

[37] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[38] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 2053951716679679.

[39] John Monahan and Jennifer L Skeem. 2016. Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology* 12 (2016), 489–513.

[40] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[41] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 39–48.

[42] Gualtiero Piccinini. 2015. *Physical computation: A mechanistic account*. Oxford University Press, New York.

[43] Eric Potash, Joe Brew, Alexander Loewi, Subhabrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, Emile Jorgenson, Raed Mansour, and Rayid Ghani. 2015. Predictive modeling for public health: Preventing childhood lead poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2039–2047.

[44] J Brendan Ritchie. 2019. The content of Marr's information-processing framework. *Philosophical Psychology* 32, 7 (2019), 1078–1099.

[45] Jean-Paul Sartre. 1998. Paris under the Occupation. *Sartre Studies International* 4, 2 (1998), 1–15.

[46] Ervin Staub. 2014. Obeying, joining, following, resisting, and other processes in the Milgram studies, and in the Holocaust and other genocides: Situations, personality, and bystanders. *Journal of Social Issues* 70, 3 (2014), 501–514.

[47] Peter Sterling and Simon Laughlin. 2015. *Principles of neural design*. MIT Press, Cambridge.

[48] Jim Torresen. 2018. A review of future and ethical perspectives of robotics and AI. *Frontiers in Robotics and AI* 4 (2018), 75.

[49] Shannon Vallor. 2015. Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology* 28, 1 (2015), 107–124.

[50] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[51] Ben Wagner. 2018. Ethics as an escape from regulation: From ethics-washing to ethics-shopping. In *Being Profiled: Cogitas Ergo Sum 10 Years of 'profiling the European citizen'*, Erme Bayamlioglu, Baraliuc Irina, Liisa Janssens, and Mireille Hilderbrandt (Eds.). Amsterdam University Press, Amsterdam.

[52] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave. 2019. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. *London: Nuffield Foundation* (2019).

[53] Niklaus Wirth. 1976. *Algorithms+ Data Structures= Programs*. Prentice Hall.

[54] Stephen Yablo. 1992. Mental Causation. *The Philosophical Review* 101, 2 (1992), 245–280.