RIVISTA INTERNAZIONALE DI FILOSOFIA E PSICOLOGIA
DOI: 10.4453/rifp.2020.0014
ISSN 2039-4667; E-ISSN 2239-2629
Vol. 11 (2020), n. 2, pp. 223-237

RICERCHE

# Driven towards a moral crash

Antoni Lorente[(α)]

Ricevuto: 12 ottobre 2019; accettato: 1 maggio 2020

**Abstract** Accidents will survive the outbreak of driverless cars, but their moral implications will suffer substantial changes. The decision made today by a human in a fraction of a second will eventually be replaced by an algorithm subject to moral scrutiny. This not only raises the question of how the algorithm should work, or whether alternatives solutions are indeed comparable, but also changes the essence of the problem: from ascertaining liability to defining desired outcomes. In this paper, I first contest two possible approaches to resolving the driverless car dilemma – which I call statistical and nominal – to conclude that neither is morally sound. I then propose an alternative solution based on "time-relative equality", or always sparing younger people. This follows from seeking (i) an egalitarian solution, that is (ii) the least intrusive – a position I defend from a possible ageist critique.
KEYWORDS: Driverless Cars; Accident; Algorithm; Age-relative Equality

**Riassunto** *Condotti verso un incidente morale* – Vi saranno incidenti anche dopo la diffusione delle auto a guida autonoma, ma le loro implicazioni morali subiranno mutamenti sostanziali. Le decisioni prese oggi in una frazione di secondo dagli esseri umani saranno alla fine sostituite da un algoritmo soggetto a controllo morale. Sorge non solo il problema di come dovrebbe funzionare questo algoritmo – o se vi siano soluzioni alternative comparabili –, ma cambia anche la sostanza del problema: dall'accertamento dell'imputabilità alla definizione dell'esito auspicato. In questo lavoro discuterò due possibili approcci per risolvere il dilemma – che chiamerò "statistico" e "nominale" –, per concluderne che non sono moralmente corretti. Proporrò quindi una soluzione alternativa, basata sulla "eguaglianza tempo-relativa", il principio per cui bisogna sempre salvare il più giovane. Questo segue dal cercare (i) una soluzione egalitarista, che sia (ii) la meno intrusiva – una posizione che difenderò da una possibile critica che la vede come discriminatoria verso gli anziani.
PAROLE CHIAVE: Automobili a guida automatica; Incidenti; Algoritmo; Eguaglianza relativa all'età

## 1 Introduction

EACH TECHNOLOGICAL LEAP BRINGS WITH it new questions; the advent of driverless cars is no exception. Replacing human drivers with software raises certain moral concerns, especially when it comes to justifying reactions to an accident situation. Regardless of the present legal constraints and remaining technical challenges – which seem to progressively pale into insignificance – driverless cars will soon trigger a paradigm shift: accidents will no longer be merely a matter of liability. In this paper, I consider a specific type of setting: accident situations involving humans – passengers and pedestrians – where the driverless car can do

[(α)]London School of Economics and Political Science, Houghton Street - WC2A 2AE London (UK)
E-mail: toni.lorente.martinez@gmail.com (✉)

something, thereby determining who will die. Regardless of the criteria behind this decision, in this scenario, it is the car's actions that define the aftermath of the accident. I shall call this *The Driverless Car Problem.*

Two seemingly feasible approaches to this moral dilemma come to mind rather easily. The first evokes the idea behind *The Moral Machine Experiment*, an exercise developed by Edmond Awad and colleagues.[1] In their research, Awad and colleagues designed a set of accidents scenarios which involved several individuals characterized in terms of nine categories, including fitness and social status, among others. The idea behind this paradigm was to survey opinions about which individuals a driverless car should spare in each accident scenario, so as to infer the preferences of the majority. I will call this a *statistical approach*. A second option is to rank individuals using a score that reflects their deeds, much like the Chinese Social Credit System – a system that evaluates individuals' behaviors as either positive or negative and translates them into a score.[2] In the *Driverless Car Problem*, this approach would demand that priority be given to those with higher scores. Despite the obvious technical limitations such an approach entails, it is also worthwhile exploring whether it could provide a fair solution to the dilemma before us. I will call this an *individual approach*.

However, I will defend neither of these approaches and instead propose a third option based on age-relative equality. The paper is structured as follows. Section 2 offers a two-fold critique of the statistical approach that focuses on the role of the majority and categories. Section 3 delves into the practical shortcomings of the individual approach. Last, In Section 4, I develop my solution to then finish with some concluding remarks.

## 2 The statistical approach

### *2.1 What the people think*

Whether a fair theory of justice can, or even must be derived from public opinion is controversial. David Miller's article *What the people think* triggered an intense critical exchange that serves to highlight some of the tensions engendered by a statistical approach to the *Driverless Car Problem*. After discussing his argument, I will go on to challenge the relevance of categories and majorities in seeking a moral solution to the case under examination.

Briefly, Miller argues in favor of defining the principles that should guide just distributions in society in line with people's normative beliefs. He defends this position first to avoid a theory of justice that lacks practical force, and second to ensure this theory ends up resting on principles of justice that do not rely (solely) on the opinions of philosophers (for as he points out, philosophers should not be assumed to have any epistemological privilege).[3] However, inferring what is just from "what the people think" invites two direct critiques. On the one hand, it is practically impossible to obtain unanimous agreement from such a survey. One can hardly claim that "the people" think something, since every majority will be accompanied by a dissident minority.[4] On the other hand, there is a pressing epistemological concern: can we ever really know what "the people" think?[5] Arguing that general opinion will guarantee a just outcome may be troublesome.

One of Miller's approaches to discovering how people make judgments of fairness about society-wide distributions of resources is to present «people with a series of "vignettes" in which a hypothetical person is described – occupation, marital status, etc. – together with his or her income, and then ask the respondents how over- or underpaid they think that person is».[6] There is, however, little discussion about the assumptions on which this approach relies.

The conceptual framework for the *Moral Machine Experiment* evokes the same idea of using "vignettes" to evaluate and distinguish the moral stance of an individual in the face of an ethical dilemma, given a set of broad categories. With this, I intend to cast doubt on the

assumption that a morally acceptable ranking is feasible by means of weighting the nine factors proposed by Awad and colleagues, namely «sparing humans (versus pets), staying on course (versus swerving), sparing passengers (versus pedestrians), sparing more lives (versus fewer lives), sparing men (versus women), sparing the young (versus the elderly), sparing pedestrians who cross legally (versus jaywalking), sparing the fit (versus the less fit), and sparing those with higher social status (versus lower social status)».[7]

Both Miller and Awad and colleagues seem to accept, as the starting point for their argument, that a preference based on majority opinion (as determined from the surveyed opinion) can constitute the moral substance for resolving an ethical dilemma. The latter, despite insisting on the indicative character of their findings,[8] do so by entering into a formal syllogistic fallacy, that is, claiming that decisions *need* to be aligned with what the majority would do. The modal fallacy lies on their methodological approach, which conflates a claim about the importance of consensus with its status as a necessary condition. This modal fallacy then allows them to argue that the ethical foundations of the algorithm *must* be rooted in majority opinion. Market acceptance, in this case at least, cannot be set as a *sine qua non* condition for the outcomes of an algorithm to be morally acceptable. Doing so could potentially institutionalize discrimination. With this, I do not intend to claim that the people are wrong. Instead, I intend to highlight how, by inferring moral judgments from statistical significance, predominant beliefs that do not favor just outcomes could very easily acquire support.

The seemingly unavoidable existence of dissident minorities and the impossibility of *really knowing* "what the people think" can be understood from Andreas Busen's interpretation of what David Miller means by "what the people think". As Busen suggests, Miller makes a distinction between two types of "thinking". The first, which he discards, consists of the average of everybody's normative beliefs. Instead, Busen suggests that what Miller is implying is a second kind of "thinking": the type of «social knowledge that is (re)-produced in social practices».[9] However, establishing such knowledge is not only difficult but also dangerous: it is quite likely to stem from «currently dominant social norms, institutions, and practices that result from hierarchical power relations, ideology or the exclusion of certain people or groups».[10] This clarifies the connection I aimed to establish between majorities and dominance.

Thus, for any statistical approach to be successful, it must be able to morally justify, at least, two crucial features: first, the use of categories to generate the input data, and second, decision-making by the majority. The following subsections discuss whether a sound defense of these features is possible by discussing some of the limitations of this view.

### *2.2 On categories*

From a social psychological perspective, categories respond to a natural human tendency to streamline cognitive resources. In order to form impressions about others, we tend to classify subjects according to their similarity to a *stereotype*, or the most representative members of a category.[11] When we conform to this cognitive mechanism, we act as "cognitive misers".[12] This form of heuristics is unbeatable from an evolutionary perspective, for it is time-efficient and «clarifies and refines our perception of the world. [...] As such, categorization provides meaning, reduces uncertainty, and helps us predict social behavior, providing prescriptive norms for understanding ourselves in relation to others».[13]

Yet categorization also poses a two-fold philosophical problem: on the one hand, categories are world-forming (i.e. their own definition constrains what is described) while, on the other hand, they break a continuous spectrum into a discrete set of hermetic conceptual containers. That our vision of the world is shaped by categories is hardly questionable. However, and despite its efficiency,

this vision typically reinforces stereotype-consistent biases, especially those reflecting social distinctions, such as race or ethnicity.[14]

To "define" a set of categories implies acknowledging clear differences between two individuals. In this sense, when surveying someone's opinion about a specific topic, the person asking the questions *must* choose or define the categories that lay out the options: race, age, profession, gender... Asking about different states along these dimensions – which, henceforth, become categories – is implicitly requiring the respondents to acknowledge the existence of these very categorical differences. At this point, it becomes imperative to question whether these categories are even relevant for the study.

Some differences could be significant, let us say, for classifying humans who take a medical test. The possibility of splitting them up into different groups (e.g. male and female) might indeed be crucial to making sense of the results. But scientifically-purposed differences are instrumental, and extending their reach into the political arena can be damaging.

Race, for example, is a widely contested category. Historical strategies to fight racial oppression have shifted from «accentuating sameness (during Abolition and the Civil Rights Movement) when racists emphasized race-conscious particularism, to praising difference (during the Black Power struggle) when whites insisted on color blind universalism».[15] However, the underlying problem is not related to the evolution of the anti-racist struggle. Instead, it is rooted in the external imposition of race as a classifying unit. In short, phenotypic differences should not inform political categories, since that could very easily lead to upholding discrimination based on a not-so-clear distinction, that is in turn not so clearly relevant.[16] Thus, when pursuing a fair system of distributive justice or a solution to a moral dilemma, arguments should depart from this ambiguous construction.

On a separate account, categories are simplifications of a continuum. As cognitive shortcuts, they reduce the complexity of case-by-case analysis, and thus comprehend a range of similar but not equal subjects. Individuals present different levels of representativeness with respect to a stereotype. This variability can relate both to intragroup membership and intergroup structure, thus leading to *heterogeneous categories* (groups with high internal variability) and *homogeneous categories* (with less intra-group differences).[17] Individuals located at the edges of two contiguous categories will naturally challenge the assumption of an "essential difference", raising a *continuity problem*. Those closer to the edge of two different categories will share similar traits, underling the weakness of conceiving of clear-cut categorical differences.

With this, I want to point out that regardless of the usefulness and the extensive application of heuristics as a cognitive shortcut, social categorization as a means for resolving moral dilemmas is not justified. For this reason, I argue that the methodological approach in the *Moral Machine Experiment* is flawed: respondents choose between options set by researchers, who assume that being or not being fit, male or female, or higher in social status constitutes solid grounds for resolving a moral dilemma. But are these options actually relevant and clearly distinguished? Apparently, we cannot escape categories. But if they are to be used in ethical dilemmas, they must be relevant – and, at this point, I cannot make the claim that the factors proposed in the *Moral Machine Experiment* are clearly relevant.

### 2.3 The concept of majority

There are several accounts that claim, with different levels of emphasis, that an algorithm must be acceptable to a majority of the population in order to enjoy moral success – this resembles Miller's idea of a fair system of distributive justice. To understand, however, the relevance of a majoritarian stance, it is necessary to define what a major-

ity is – an idea that cannot be fully detached from another related idea, that of the dissonant minority.[18]

On the one hand, and from the socio-psychological standpoint, negative stereotypes are mistakenly associated with minority groups through a mechanism called *illusory correlation*. This consists of wrongly believing that two variables are correlated when there is little or no actual association. This can be explained by the notion of *shared distinctiveness*, which consists in connecting infrequent characteristics with infrequent individuals.[19] The prevalence with which negative stereotypes are associated with minority groups is then maintained by a natural tendency to stress in-group similarities and out-group differences. On the other hand, a statistical approach to majority reveals how the majority-minority duality arises naturally from the aftermath of surveying a population with different views on a given topic. As a consequence, some options will have more support than others – that is the majoritarian view. But what do majorities *mean*?

Arrow's impossibility theorem is one of the most well-known ideas in social choice theory. It explains that there is no way to transform individual ranked preferences into social or community-wide ranked preferences without violating one of the following five conditions: unrestricted domain, pareto optimality, independence of irrelevant alternatives, not imposing the social welfare function, and a non-dictatorial social welfare function.[20] Hence, we can infer that polling individual preferences with the intention of drawing clear social preferences will be impossible. As Arrow stated, «if we exclude the possibility of interpersonal comparisons of utility, then the only methods of passing from individual tastes to social preferences which will be satisfactory and which will be defined for a wide range of sets of individual orderings are either imposed or dictatorial».[21] Hence, majorities are delusional simplifications of "what the people think". To this, I shall add that the statistical account presumes a non-existent equality between individuals, without due consideration of historical contingencies and patterns of discrimination.

Thus, buttressing the solution to the *Driverless Car Problem* on a statistical majority will imply either misrepresenting a complex landscape of preferences, or violating one or more conditions of Arrow's impossibility theorem – meaning that such an approach will see its normative capacity diminished. Moreover, surveying individuals' preferences about their fellow citizens is more likely to reflect embedded relations of oppression than to provide a just solution to the dilemma. This is well illustrated by the *Moral Machine Experiment*, which reported a "statistical preference" to spare dogs before criminals[22] – a result that provides food for thought.

By knowing the opinion of the majority, one can easily make an algorithm and its rationale match the general opinion of what ought to be done and why. But it is a different thing to claim that such acceptance is necessary for it to be morally acceptable. This form of reasoning responds to the modal fallacy I highlighted earlier, a form of inferring necessity that *seems* to follow due to a cultural bias towards market-based decision-making. In sum, an accepting majority makes an algorithm more popular but does not suffice to make it morally acceptable.

## 3 An individual approach

To move away from a statistical approach means to assess each individual as an independent unit and not as a member of a group. As in China's Social Credit System, each person is thus allocated a unique score. The source of this score could be virtually anything, but a detailed description of what makes an individual scoring system fair falls outside the scope of this paper. What is remarkable about this approach is the underlying idea that individuals should be compared via a score based on their deeds. This, however, entails several problems.

Danielle Citron and Frank Pasquale, in their research paper *The Scored Society*, point to some of the essential flaws linked to the use of *predictive algorithms*, a form of the individual approach that makes predictions about future events based on historical facts. They do so by contrasting the opportunities and threats these algorithms present.[23] Through a case study of scoring systems used to calculate financial risk, they highlight the main deficits of automated scoring systems. Citron and Pasquale argue that these systems are opaque, foster arbitrariness, and have disparate impacts. However, the problem with this view is two-fold. On the one hand, arbitrary assessments and disparate impacts are not consequences, but potential features of a predictive algorithm – even more so if it is developed under a veil of opacity. On the other hand, they seem to overlook the reasonable concern that such algorithms will lead to increased threats to privacy .

However, and to digress, their analysis suggests that there is nothing intrinsically wrong with individual scoring systems. On the contrary, what they find morally reprehensible are both bad implementations and lack of oversight. Authoritarian decisions are not seen to be a hallmark of scoring systems, but the byproduct of poor implementations.

In this section, I first discuss the role of transparency in predictive algorithms. I then continue by introducing the possibility that these algorithms will become a threat to privacy. I finish with a brief account of surveillance capitalism to better understand the extent of this approach.

### *3.1 Lack of transparency*

Without transparent algorithms and control over automated systems, «societies are destined to continue to reinforce patterns of entrenched privilege and disadvantage, widening gaps between rich and poor, and perpetuation of disadvantage».[24] Transparency allows individuals and experts to identify features that could end up triggering arbitrary decisions.

Far from being sufficient, access to the source code of algorithms is nevertheless important; it provides an opportunity to challenge their logic and detect possible sources of bias. But transparency *per se* does not guarantee a just outcome. Its utility is instrumental, acting like a bridge of confidence between the subjects under evaluation and the system – which, in turn, must run on fair criteria and be implemented systematically.

Fair criteria must be the essence of any decision tree that aspires to be just. As Citron and Pasquale highlight, decisions-by-algorithm must avoid *categorization by correlation* – that is classifying people based on categories such as race, gender, or sexual orientation.[25] The very same critique was applied to statistical approaches in the previous section, with a formal nuance that constitutes, in fact, a major conceptual difference: in statistical approaches, categorization [of different kinds] is inevitable, whereas in individual approaches correlative categories can be more easily avoided.

Moreover, it is necessary to ensure that principles are implemented in a systematic way. The convergence of fair criteria and their systematic application would ensure that people were treated in a just manner. The danger of arbitrariness and discrimination would then disappear, making it easier for the algorithms to go from the «illusion of precision and reliability» to actually becoming accurate and fair.[26]

One aspect of this question is whether the algorithm risks being authoritarian if it remains in a black box, even if the principles behind it are fair and its implementation systematic. What role does publicity play? For Andrew Mason, the reason for publicity – or transparency – «derives from two sources. [...] The first source of its appeal is the idea that justice itself is better promoted by publicly checkable rules; the second source is the idea that publicity tends to promote stability».[27] However, publicity is not a necessary feature to prevent a scoring system from becoming authoritarian. This could be achieved

by "technological due process" – or the idea that algorithms should «live up to some standard of review and revision to ensure their fairness and accuracy».[28] The second aspect of this question is whether making an algorithm transparent suffices to make the system fair. This, I believe, I have already answered. But for the sake of clarity, transparency itself does not make an algorithm fair, it just increases its chances of being fair.

### 3.2 A threat to privacy

Designing fair, systematic, and transparent algorithms is not enough to wash away the threat of authoritarianism. That is because even when a system presents such values, the possibility of it becoming too invasive remains. Authoritarianism can manifest here in two ways. The first, I have already covered: it can l unjust decisions based on unfair criteria, and/or a bad implementation, all sustained by lack of transparency. The second concerns overreach.

Granting access to our personal data is something we constantly do – to the government, to our service providers, to our friends... The main difference between each case is the form and extent of our consent. The internet is one of the most shady cases, since we often consent without wanting to or knowing that we have done so. We condone "*unconscionable contracts*" – namely exploitative contracts between data subjects and service providers buttressed on the inability to negotiate[29] – sometimes because we do not even know that we are bound by one, and at other times because there seems to be no better alternative.

While we are online, we contribute to images of ourselves that are formed via the internet and social media. Lori Andrews outlines how every click on a social network is an earmark that contributes to the assembly of your "cyber-self" – a virtual representation of who you are. In turn, every trace and observation is used to influence decisions and opportunities that concern you – i.e. your access to mortgages, jobs, and discounts amongst others.[30] In an attempt to control this, Europe has developed the so-called *General Data Protection Regulation*, which intends to look after European citizens by moving away from a paradigm of abusive relations between companies and users and turning this relationship into one based on affirmative consent, that ensures easier access to our own personal data and grants the right to be forgotten.[31]

All of this leads to the virtual self being most commonly invoked when an individual approach to the *Driverless Car Problem* is suggested. However, I believe there is a substantial difference between the profiling we are subjected to on the internet and an individual approach to the *Driverless Car Problem*. Whilst the first aims to yield a benefit from the use of personal information, the second aims at resolving what we consider a moral dilemma.

That obvious – and not so obvious – differences between individuals exist is hardly debatable. In practice, however, the equilibrium between an egalitarian ethos and our conception of our own personal exceptionality is fragile. My argument does not challenge the existence of differences between individuals. But whether these differences justify an unequal score and, consequently, unequal treatment seems less clear. And that in turn demands that we ask if such differences are even relevant to the stated problem – a question I address in Section 4.

### 3.3 Surveillance capitalism

Individual approaches are commonly associated with "flawed democracies" or other political systems that do not respect individual liberties. But corporations in western democracies regularly violate certain forms of individual rights through dataveillance practices or the monitorization of online personal data and metadata.

Based on the ideals of human rights, Jonathan Cinnamon argues that privacy provides

the grounds for freedom of speech and association, thus constituting «a unifying narrative in democratic societies and a key concept invoked to challenge escalating practices of dataveillance». However, dataveillance is a practice mainly conducted by corporations that seek to make a profit from personal information. The problem with scoring systems is largely related to their lack of regulation and transparency. Using the example of the Chinese Social Credit System, Cinnamon argues that the Chinese are facing a future in which their «identity and social status will become increasingly externally shaped [...] rather than intersubjectively through equitable social relations of recognition».

Surveillance capitalism is the capitalization by companies of their dataveillance practices. The term was coined by Shoshana Zuboff to describe several processes by which corpo-rations commodify personal information, which has fostered the development of big data analytics. The cornerstone of big data analytics, in turn, is that every actor, event, or transaction can be made visible. However, the capacity to monitorize and use this information is held by very few companies. Additionally, the economics of personal data is first based on dubious business ethics and then sustained by undue technical obstacles and unsatisfactory laws and regulations.

Against the corporate use of personal data, Cinnamon uses Nancy Fraser's theory of ab-normal justice. This framework stems from three obstacles to parity of participation: maldis-tribution (not having the ability to participate equally in social life due to lack of resources), misrecognition (the inability to shape one's identity due to institutionalized hierarchies) and misrepresentation (the inability to control one's own representation). Briefly, according to this view, when a process prevents individuals from enjoying parity of participation, it becomes unjust.

Although this framework serves to characterize the abnormal justice derived from the accumulation of personal data by corporations, it fails to raise an exhaustive critique of scoring systems. That is because theoretical scoring systems have multiple positive aspects. It is in their implementation that the fragile equilibrium is potentially broken, jeopardizing the privacy of those subjected to them. And the danger of an authoritarian use of a database with detailed profiles of all citizens seems serious enough danger to look for a solution elsewhere.

## 4 Restating the problem

In this section, I begin by drawing an analogy with the continuum model from social psychology. I then explore why the *Driverless Car Problem* entails a hard choice, to finish with a defense of a solution to the dilemma based on *age-relative equality*.

### *4.1 Continuum model*

The solutions hitherto considered evoke the extremes of what Susan Fiske and Steven Neuberg called the *continuum model of impression formation*. Their model intends to explain how individuals form impressions of others. One extremity corresponds to *category-based* (or heuristic) *processing* and the other to *attribute-based* (or systematic) *processing*. This model explains how we naturally try to fit strangers into categories to save cognitive resources, but also how, when doing so becomes problematic because "the other" does not fully match the stereotype, we shift towards an individuated analysis. This transition from heuristics to systematic processing is known as *decategorization* and allows us to classify subjects as individuals and not as group members. This, in turn, serves to eliminate category-based biases.[32] A solution to the *Driverless Car Problem* based on an individual approach would thus foster decategorization.

The nature of the problems that individual and statistical solutions face is substantially different: on the one hand, statistical approaches are doomed to fall into correlative categories, making it hard to think of a *Moral*

*Machine Experiment* that is in fact morally acceptable. On the other hand, individual approaches seem to be constrained by sufficiently serious practical difficulties that push us to explore alternative responses (since the possibility of individual scoring systems degrading into authoritarian instruments is too high a risk to bear). Neither of these solutions seems good enough.

To complete the analogy, what can also be inferred from the continuum model is that the extremities are *only* two points within a greater range of available options. Therefore, we need not be satisfied with one of the two without first exploring other options. Both approaches – the individual and the statistical – embody the idea that individuals can be differentiated and compared. What makes each point of the continuum different is the amount and nature of features known about the individuals in each case. Therefore, defining the threshold for what constitutes sufficient information to make a *morally acceptable distinction* in terms of both quantity and quality will also define the nature of the solution, pinpointing it along "the continuum".

Finally, and with regard to a possible scenario where more than one solution seems feasible, I shall defend the Minimum Invasion Principle: If there are two or more solutions, one requires less information than the others and all seem morally sound, the less invasive one should be chosen. This I will defend shortly.

### 4.2 Trapped in a hard choice?

*The Driverless Car Problem* is a *targeting problem*.[33] It demands we first define the possible targets and then design and implement the criteria to choose between them. But to resolve whose life ought to be spared seems to be a hard choice. In this regard, however, I hope that by better understanding just what makes this a hard decision, we can come up with a fair solution. To do so, I use Ruth Chang's work on hard choices.[34]

Let us consider a rational agent facing a *hard choice*, which is in turn defined as a situation in which the agent must decide between two alternatives when «one alternative is better in some relevant aspects, the other is better in other relevant aspects and yet neither seems to be at least as good as the other overall in all relevant aspects».[35] In hard choices, the agent's reasons to make one choice or the other run out because the alternatives don't match the trichotomy "better than", "worse than", and "equally good". Thus, it is neither ignorance, nor incommensurability, nor incomparability that makes the choice hard, but the fact that alternatives are *on a par*.[36]

Should the *Driverless Car Problem* turn out to be a problem that has emerged from ignorance, incommensurability, or incomparability, the decision would then be easier than we thought. However, asking whose life – Mary's or James's – is more valuable, seems in fact a problem of parity and, therefore, a hard choice. But what if we could restate the problem in different terms? We then might be able to avoid a choice between options on a par, escaping what would otherwise become a moral trap. More on this in due course.

Chang differentiates between the choices in which one has *first-personal authority*, when «your judgement [...] determines the truth of the matter, give or take a margin of error»,[37] and the ones in which that is not the case. The *Driverless Car Problem*, if any, would fit into the second kind of choice, for "The Decider" [i.e. the algorithm or, ultimately, the programmer] is not *directly involved* in the situation – the criteria behind the decision-making process are set beforehand. One way to regain first-personal authority could be by compelling the users to choose a decision tree they liked before starting the journey. Yet the cost of doing so would be unbearable, for the passengers could opt for one that protected them over the others in all cases, granting them undue power that the other parties would not have.

She then makes a case for the relevance of *practical certainty* rather than knowledge. Ig-

norance or uncertainty of certain data that is relevant to making a choice is not the true problem, but rather the fact that «none of the usual trichotomy of relations [better than, equally good, and worse than] holds for the alternatives».[38] It is the way alternatives relate to each other and not us being uncertain that makes a choice hard.

If the problem is lack of information, the solution will involve using more detailed profiles of the subjects affected. But that will not suffice, for once you perfectly know Mary and James, you will realize that you have two alternatives that are, *in respect to what matters* (or V, borrowing Chang's notation) very different, «but neither is better than the other and nor are they equally good. They are on a par».[39] So what criteria V should the alternatives be compared on?

The manner in which alternatives relate to each other in the *Driverless Car Problem* depends on how V is defined. If V was, let us say, "the value of someone's life", I fear Mary and James would be on a par, for it can be argued *prima facie* that one is not better than the other (their evaluative difference is unbiased) but they are not equal (the magnitude of such evaluative difference is nonzero) – this, despite the fact that finding an uncontroversial definition of "the value of someone's life" seems to be a task beyond the scope of this exercise. Unfortunately, the solution proposed by Chang when we have to make choices with alternatives "on a par" relies on *will-based choices*, which involves creating our own will-based reasons to justify a decision in a hard choice,[40] creating a theoretical cul-de-sac for the case discussed here.

The hard choice here is the one that the software engineer designing the decision tree of the algorithm will face at her desk – or we face from the comfort of our armchairs. We no longer have autonomous agents "making" their own reasons to decide in a situation that involves them directly, but a design team that must decide how decisions that do not affect them *should* be made (this "should" reflects the moral component of the problem). Consequently, arguing for a process of making our own will-based reasons – in Chang's terms – does not seem to offer a solution to the *Driverless Car Problem*.

But V need not be "the value of someone's life". There is a whole range of possible criteria, each one bearing different pros and cons. In light of this, and consistent with the *Minimum Invasion Principle*, I advocate for the least intrusive amongst all the morally acceptable options. That is, one for which the amount of information required about the subjects involved is as non-invasive as possible but that minimizes, at the same time, any possible moral recriminations in the aftermath of the decision. The motivation behind this principle is to ensure, to the extent possible, that the privacy and individual rights of those involved in this type of accidents will be preserved. The discussion should then gravitate around how to define the criteria "V" to which alternatives are compared, making differentiation between subjects a consequence of the nature of V. On this account, once the criterion is established, the amount and precision of the information required to differentiate between individuals is a corollary. By implementing a morally acceptable criterion that allows for making a decision based on the classic trichotomy [better than, worse than, and equally good], we might find a way out of this maze. But what criterion could we use?

### 4.3 Time for equality

Patrick Lin, in *Why ethics matters for autonomous cars*, remarks on the importance of ethics in the *Driverless Car Problem* through an example. Imagine an autonomous car that must make a choice: «it must either swerve left and strike an eight-year old girl, or swerve right and strike an 80-year-old grandmother»;[41] inaction, however, will result in the killing of both individuals. Lin suggests that, in this case, many would claim "the lesser evil" would be swerving to kill the grandmother. Mainly for two reasons: «that the girl still has

her entire life in front of her [...] [and that] the little girl is a moral innocent».[42]

But Lin claims that neither option is moral. The main point of his thesis is that age is irrelevant in this scenario. To support this, he contrasts it with another example where age seems to be more determining (i.e. rejecting adult actors for a child's character in a movie).[43]

He concludes by stating that «a reason to discriminate does not necessarily justify that discrimination, since some reasons may be illegitimate».[44] «Discriminating on the basis of age in our crash scenario would *seem* to be the same evil as discriminating on the basis of race, religion, gender, disability, national origin, and so on, even if we can invent reasons to prefer one such group over another».[45] However, and although it does *seem* to be the same evil, it is not. For even though regulators intend to or have already prohibited making age-based decisions,[46] I believe there is a strong case for using age as the criterion.

Time constitutes an *organic unity* (continuing with Ruth Chang's terminology). That is, time's quantity and quality are not «independent determinants of the evaluative difference in V-ness [the comparison criterion] between two items, [...] they are interdependent».[47] I believe, however, that "time" entails even more than that. Time is the non-reversible context in which one's life unfolds: who we are is, somehow, defined by how we spend our time. And age, as a time-dependent variable, is different from other [more or totally] static variables such as race, gender, or national origin. Against Lin's view, I will defend the argument that sparing younger individuals constitutes a morally acceptable solution to the *Driverless Car Problem*.

First, age, due to its time-dependent nature, is equalizing. Unlike race, gender, sexual orientation, and other categories of the kind, everyone's age changes equally. Thus, an individual's age will not be a static number but a value that increases as time passes – and it will increase at the same rate for everyone. If this criterion is implemented, the car will swerve to protect younger individuals. Extending this logic, it is easy to see that a newborn baby would have one hundred percent probability of being the youngest subject involved in the accident. Because of that, she will have the greatest chance of being spared. However, as her age advances, her probability of being the youngest involved will decrease, therefore increasing her chances of being targeted.

Given any individual, if we put together all the values for the probability of being the youngest person involved in a driverless car accident that she has held throughout her life, we obtain her *life-distribution of probability*. At a particular moment, everyone holds a different individual value, which depends on each one's age at that time. Yet the overall life-distribution is the same for everyone. It starts from its maximum in terms of the probability of being spared and then decreases over time.

Thus, the older person, now being targeted by the car, was once in the position of the younger person, who is now spared. Yet the spared one will be, at some point in the future, in the position of the one who is being run over. Unlike the other features that Lin mentions, or the categories that the *Moral Machine Experiment* suggests, age provides room for equality, for every individual is bound to have a similar life-distribution of probability of being saved. This I shall call *age-relative equality*.

Individuals are given preference according to their age, and that could be interpreted as *ageism*. However, this form of differentiation departs radically from what Robert Neil Butler first described as ageism. Saving the young is not proposed here out of a system of beliefs and prejudices held [by the middle-aged] against the old or the young, that reflects «a personal revulsion to and distaste for growing old, disease, disability; and fear of powerlessness, "uselessness", and death».[48] On the contrary, the reason for sparing the young(er) is to ensure that everyone is treated equally. This solution is not a defense of the "lesser evil solution", buttressed on the

young having their whole life ahead of them and being "moral innocents", as Lin suggests.[49] Neither does it stems from an alleged value intrinsically linked to youth, that grants younger people some sort of metaphysical privilege. The difference proposed here rests on the idea that everyone throughout their lives should have the same distribution of probability of being saved. This difference, however, becomes less significant when the ages of potential targets are relatively similar. Should age-relative equality suffice to make a decision between a fifteen- and a seventeen-year-old? What if the individuals are just a few months apart? These cases highlight some of the tensions of the approach here presented. If one is not willing to bite this bullet, age-relative equality will need further conditions to address small age gaps. But the possibility of a nuanced approach will demand more invasive profiles – and that in turn redirects us to a statistical approach, the downsides of which have been discussed at length.

Second, I have defended the position that majority acceptance is not a necessary condition for a decision to be fair (see Section 2). However, and from a practical point of view, the support of an accepting majority allows for easier implementation of the solution. In this sense, Awad and colleagues found that the majority of respondents to their experiment preferred sparing "the stroller", "the girl" and "the boy".[50] Thus, the majoritarian intuition is consistent with the principle. This convergence, however, is deductively independent from my argument, for the reasons behind the appeal to equality – although counting with an accepting majority surely represents a practical benefit.

Finally, age seems to imply a rather small sacrifice in terms of disclosure; it is non-invasive, in the sense that age is not a private datum that would put the construction of one's persona at stake. Moreover, and from an instrumentalist perspective, it is fairly easy to implement: although the solutions to the *Driverless Car Problem* are limited by technology as well, it is plausible to imagine a way for the car to know precisely what age anyone involved in an accident may have: via small devices, precise facial recognition, or a more futuristic solution. I leave any possible moral implications related to the mechanisms used to harvest the data open for further research.

## 5 Concluding remarks

The aim of this exercise has been to find a moral solution to the *Driverless Car Problem* by contesting two possible approaches: a statistical approach, similar to that of the *Moral Machine Experiment*; and an individual approach inspired by the Chinese social credit system.

I first criticized statistical approaches. Their use of correlative categories and majority-based decision-making procedures serves to institutionalize the patterns of discrimination embedded in society. I have also argued that statistical approaches are over-reductive and entail a degree of uncertainty that seems to challenge the validity of any solution to a moral dilemma that stems from any alleged majority.

On the other hand, I have shown how individual approaches can very easily become authoritarian. To avoid this outcome, I have suggested that such approaches should bring together fair principles, systematic implementation, and transparency. When these principles are observed, individual scoring systems come with multiple benefits. But since a bad implementation could pose a serious threat to privacy or allow for thee wrongful use of detailed information about citizens, I have argued that a solution that withstands moral scrutiny will not resemble either of these two alternatives. Other options could have been considered: tossing a coin, random choice, or saving the driver no matter what, *inter alia*. However, neither chance nor undue priority seem to make a strong enough case to leave off seeking a better alternative. My intention has not been to fully develop the schemata for a morally good

algorithm. On the contrary, I intended to challenge some of the assumptions that are being made in order to predict what the essence of a fair solution to the *Driverless Car Problem* would look like.

*The Driverless Car Problem* involves a hard choice. The value of the life of the individuals involved in a car accident is not equal, but neither is one individual better than the other: their evaluative difference is unbiased but has nonzero magnitude. Therefore, they are *on a par*. However, we can change the nature of the problem and bring the solution back to the domain of what Ruth Chang calls the classic trichotomy – "better than", "worse than" and "equally good". To do so, we must discard "the value of someone's life" as the evaluative criterion and look for another one. The challenge then has been to find a criterion that can be morally justified, because what makes this choice hard is how the alternatives relate to each other and that, in turn, is determined by the criteria used to evaluate and compare the options.

I have defended the argument that basing the algorithm on an *age-relative egalitarian* criterion would provide a morally sound solution to the *Driverless Car Problem.* Giving priority to younger individuals would imply that everyone had the same life-distribution of probability to survive an accident situation. Additionally, the amount of data required to implement this approach is insignificant, making it more feasible from a practical perspective.

On a separate account, I have responded to a possible critique that might accuse this view of ageism. The main counterargument is that the differentiation proposed is a means to ensure equality. It does not reflect discrimination against the elderly grounded on age itself or on features intrinsic to advanced stages of life, but is instead based on an intention to ensure equal treatment for everyone.

With the advent of driverless cars, the consequences of an accident will no longer exclusively concern liability problems – ethics will play a crucial role. And with this paper I have tried to develop a morally acceptable solution to the *Driverless Car Problem*: one based on an egalitarian treatment of the individuals involved. If I have succeeded, a car can swerve without moral reproach. If not, we may find ourselves driven towards a moral crash.

## Notes

[1] Cf. E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.F. Bonnefon, I. Rahwan, *The Moral Machine Experiment*, in: «Nature», vol. DLXIII, n. 7729, 2018, p. 59-64.
[2] Cf. R. Botsman, *Who can you trust?: How technology brought us together – and why it could drive us apart*, Penguin, London 2017.
[3] A. Baderin, A. Busen, T. Schramme, L. Ulaş, D. Miller, *Who cares what the people think? Revisiting David Miller's approach to theorising about justice*, in: «Contemporary Political Theory», vol. XVII, n. 1, 2018, pp. 69-104, here p. 78.
[4] *Ibid*., pp. 91-92
[5] *Ibid*., p. 85.
[6] D. Miller, *Distributive justice: What the people think*, in: «Ethics», vol. CII, n. 3, 1992, pp. 555-593, here p. 565.
[7] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.F. Bonnefon, I. Rahwan, *The Moral Machine Experiment*, cit., p. 60.
[8] *Ibid*., p. 59.
[9] A. Baderin, A. Busen, T. Schramme, L. Ulaş, D. Miller, *Who cares what the people think?*, cit., p. 81.
[10] *Ibid*., p. 83.
[11] Cf. L.W. Barsalou, *Deriving categories to achieving goals*, in: M.I. Posner (ed.), *The psychology of learning and motivation*, Academic Press, New York 1991, pp. 1-64.
[12] Cf. S.T. Fiske, S.E. Taylor, *Social cognition,* McGraw-Hill, New York 1991.
[13] R.J. Crisp, R.N. Turner, *Essential social psychology*, Sage, London 2010, p. 75.
[14] *Ibid*., p. 79.
[15] R. Delgado, J. Stefancic, *Critical race theory: An annotated bibliography 1993, a year of transition*, in: «University of Colorado Law Review», n. LXVI, n. 1, 1994, 159-193, here p. 167, see also p. 168, 174 and 182.
[16] For a revealing experience that challenges the traditional dichotomic conception of race, I suggest taking a look at Angélica Daas' *Humanae project.*

[17] R.J. CRISP, R.N. TURNER, *Essential social psychology*, cit., p. 74.
[18] A. BADERIN, A. BUSEN, T. SCHRAMME, L. ULAŞ, D. MILLER, *Who cares what the people think?*, cit., pp. 91-92.
[19] R.J. CRISP, R.N. TURNER, *Essential social psychology*, cit., p. 65.
[20] Cf. K.J. ARROW, *A difficulty in the concept of social welfare*, in: «Journal of Political Economy», vol. LVIII, n. 4, 1950, pp. 328-346.
[21] *Ibid*., p. 342.
[22] E. AWAD, S. DSOUZA, R. KIM, J. SCHULZ, J. HENRICH, A. SHARIFF, J.F. BONNEFON, I. RAHWAN, *The Moral Machine Experiment*, cit., p. 61.
[23] Cf. D.K. CITRON, F. PASQUALE, *The scored society: Due process for automated* predictions, in: «Washington Law Review», vol. LXXXIX, n. 1, 2014, pp. 1-33, especially pp. 10-16.
[24] J. WOLFF, A. DE-SHALIT, *Disadvantage*, Oxford University Press, Oxford 2007, p. 186.
[25] D.K. CITRON, F. PASQUALE, *The scored society*, cit., p. 24
[26] *Ibid*., p. 33.
[27] A. MASON, *Just constraints*, in: «British Journal of Political Science», vol. XXXIV, n. 2, 2004, pp. 251-268, here p. 264
[28] D. K. CITRON, F. PASQUALE, *The scored society*, cit., p.19
[29] Cf. S.E. PEACOCK, *How web tracking changes user agency in the age of Big Data: The used user*, in: «Big Data & Society», vol. I, n. 2, 2014, pp. 1-11; see also J. CINNAMON, *Social injustice in surveillance capitalism*, in: «Surveillance & Society», vol. XV, n. 5, 2017, pp. 609-625, especially p. 610.
[30] L.B. ANDREWS, *I know who you are and I saw what you did: Social networks and the death of privacy*, Free Press, New York/London 2012, especially pp. 19-20.
[31] Cf. G.D.P.R., *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46*, in: «Official Journal of the European Union», vol. LIX, n. 295, 2016, pp. 1-88.
[32] R.J. CRISP, R.N. TURNER, *Essential social psychology*, cit., p. 85.
[33] P. LIN, *Why ethics matters for autonomous cars*, in: M. MAURER, J. C. GERDES, B. LENZ, H. WINNER (eds.) *Autonomous driving,* Springer, Berlin/Heidelberg 2016, pp. 69-85, here p. 72.
[34] Cf. R. CHANG, *Hard choices*, in: «Journal of the American Philosophical Association», vol. III, n. 1, 2017, pp. 1-21.
[35] *Ibid*., p. 1.
[36] *Ibid*., p. 16.
[37] *Ibid*., p. 3.
[38] *Ibid*., p. 5.
[39] *Ibid*., p. 15.
[40] *Ibid*., p. 20.
[41] P. LIN, *Why ethics matters for autonomous cars*, cit., p. 70.
[42] *Ibidem*.
[43] *Ibidem*.
[44] *Ibidem*.
[45] *Ibid*., p. 71 - emphasis added.
[46] See Germany, for example. Cf. E. AWAD, S. DSOUZA, R. KIM, J. SCHULZ, J. HENRICH, A. SHARIFF, J.F. BONNEFON, I. RAHWAN, *The Moral Machine experiment*, cit., p. 60; P. LIN, *Why ethics matters for autonomous cars*, cit., p. 70.
[47] R. CHANG, *Hard choices*, cit., p. 13.
[48] R.N. BUTLER, *Age-ism: Another form of bigotry*, in: «The Gerontologist», vol. IX, n. 4, 1969, pp. 243-246, here p. 243.
[49] P. LIN, *Why ethics matters for autonomous cars*, cit., p. 70.
[50] Cf. E. AWAD, S. DSOUZA, R. KIM, J. SCHULZ, J. HENRICH, A. SHARIFF, J.F. BONNEFON, I. RAHWAN, *The Moral Machine experiment*, cit., p. 61.

## References


ANDREWS, L.B. (2012). *I know who you are and I saw what you did: Social networks and the death of privacy*, Free Press, New York/London.

ARROW, K.J. (1950). *A difficulty in the concept of social welfare*. In: «Journal of Political Economy», vol. LVIII, n. 4, pp. 328-346.

AWAD, E., DSOUZA, S., KIM, R., SCHULZ, J., HENRICH, J., SHARIFF, A., BONNEFON, J.F., RAHWAN, I. (2018). *The Moral Machine Experiment*. In: «Nature», vol. DLXIII, n. 7729, p. 59-64.

BADERIN, A., BUSEN, A., SCHRAMME, T., ULAŞ, L., MILLER, D. (2018). *Who cares what the people think? Revisiting David Miller's approach to theorising about justice*. In: «Contemporary Political Theory», vol. XVII, n. 1, pp. 69-104.

BARSALOU, L.W. (1991). *Deriving categories to achieving goals*. In: M.I. POSNER (ed.), *The psychology of learning and motivation*, Academic Press, New York, pp. 1-64.

BOTSMAN, R. (2017). *Who can you trust?: How technology brought us together – and why it could drive us apart*, Penguin, London.

BUTLER, R.N. (1969). *Age-ism: Another form of bigotry*. In: «The Gerontologist», vol. IX, n. 4, pp. 243-246.

CHANG, R. (2017). *Hard choices*. In: «Journal of the American Philosophical Association», vol. III, n. 1, pp. 1-21.

CINNAMON, J. (2017). *Social injustice in surveillance capitalism*. In: «Surveillance & Society», vol. XV, n. 5, pp. 609-625.

CITRON, D.K., PASQUALE, F. (2014). *The scored society: Due process for automated* predictions. In: «Washington Law Review», vol. LXXXIX, n. 1, pp. 1-33.

CRISP, R.J., TURNER, R.N. (2010). *Essential social psychology*, Sage, London.

DELGADO, R., STEFANCIC, J. (1994). *Critical race theory: An annotated bibliography 1993, a year of transition*. In: «University of Colorado Law Review», n. LXVI, n. 1, 159-193.

FISKE, S.T., TAYLOR, S.E. (1991). *Social cognition,* McGraw-Hill, New York.

G.D.P.R., *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46*. In: «Official Journal of the European Union», vol. LIX, n. 295, 2016, pp. 1-88.

LIN, P. (2016). *Why ethics matters for autonomous cars*. In: M. MAURER, J. C. GERDES, B. LENZ, H. WINNER (eds.) *Autonomous driving,* Springer, Berlin/Heidelberg, pp. 69-85.

MASON, A. (2004). *Just constraints*. In: «British Journal of Political Science», vol. XXXIV, n. 2, pp. 251-268.

MILLER, D. (1992). *Distributive justice: What the people think*. In: «Ethics», vol. CII, n. 3, pp. 555-593.

PEACOCK, S.E. (2014). *How web tracking changes user agency in the age of Big Data: The used user*. In: «Big Data & Society», vol. I, n. 2, pp. 1-11.

WOLFF, J., DE-SHALIT, A. (2007). *Disadvantage*, Oxford University Press, Oxford.