

Fair navigation planning: a resource for characterizing and designing fairness in mobile robots

Martim Brandão^{a,*}, Marina Jirtoka^a, Helena Webb^a, Paul Luff^b

^a*University of Oxford, Oxford, UK*

^b*King's Business School, King's College London, London, UK*

Abstract

In recent years, the development and deployment of autonomous systems such as mobile robots have been increasingly common. Investigating and implementing ethical considerations such as fairness in autonomous systems is an important problem that is receiving increased attention, both because of recent findings of their potential undesired impacts and a related surge in ethical principles and guidelines. In this paper we take a new approach to considering fairness in the design of autonomous systems: we examine fairness by obtaining formal definitions, applying them to a system, and simulating system deployment in order to anticipate challenges. We undertake this analysis in the context of the particular technical problem of robot navigation. We start by showing that there is a fairness dimension to robot navigation, and we then collect and translate several formal definitions of distributive justice into the navigation planning domain. We use a walkthrough example of a rescue robot to bring out design choices and issues that arise during the development of a fair system. We discuss indirect discrimination, fairness-efficiency trade-offs, the existence of counter-productive fairness definitions, privacy and other issues. Finally, we elaborate on important aspects of a research agenda and reflect on the adequacy of our methodology in this paper as a general approach to responsible innovation in autonomous systems.

*Corresponding author

Email address: `martim@robots.oc.ac.uk` (Martim Brandão)

Keywords: motion planning, robot navigation, algorithmic fairness, ethics, responsible innovation

1. Introduction

In recent years there has been a proliferation of research concerned with the ethics of autonomous systems and artificial intelligence, sparked by investigations of the ethical dimension to many of our seemingly-neutral digital, transportation, robotic and other technologies [1, 2, 3]. This concern has led to greater pressure on developers to innovate responsibly, as well as to the development of a great number of guidelines and principles for ethical development. For example, an informal survey at the end of 2017 found that a total of 10 different sets of ethical principles had been proposed by December 2017, seven of which appeared in 2017 [4], and the number keeps growing until today [5]. Such guidelines and principles are helpful in providing a framework for researchers and practitioners. However, they are limited in terms of supporting researchers and practitioners to actually implement and satisfy the principles in practice. For example, it is often not clear how “fairness” or “beneficence” principles are relevant to a new technology, or how to respect the principles in practice.

Our approach in this paper is to unpack the concept of an ethical concern by collecting definitions from the technical and philosophical landscapes and then applying them to a technical problem. In particular we focus on fairness and what seems to be a rather mundane technical problem—robot navigation. We build a resource of formal definitions of fairness in this context, as well as a set of design options related to fairness in navigation. Together with simulations of deployment outcomes (i.e. inequalities in access to the robot) this resource can be used to guide responsible innovation of the technology in question and to better ground discussions between stakeholders at the design stage. We then reflect on the adequacy of the approach as a general method to use in the early stages of the design of autonomous systems for responsible innovation with ethical considerations.

In this paper we guide the reader through a walkthrough example of a robot navigation application to bring out the fairness dimension of the system, the fairness-related design choices, and to finally establish a research agenda for
30 responsible innovation in autonomous systems with respect to fairness.

Our contributions are the following:

1. We show the fairness dimension of robot navigation, using a walk-through example of a rescue robot to bring out concerns and contrast to other
35 robot applications.
2. We build a resource of formal definitions and design choices related to fairness in the context of robot navigation.
3. We use a new methodology for responsible innovation based on such a resource and the simulation of technology-deployment outcomes, which
40 we argue should make it easier to ground discussions with stakeholders during the early stages of design.

The paper is organized as follows. In Section 2 we overview relevant concepts of responsible innovation, algorithmic fairness, discrimination, distributive justice, and issues of context and trade-off in fairness. Then, in Section 3, we use
45 a walkthrough robot navigation example to argue for the existence of a fairness dimension to navigation planning. In Section 4 we outline multiple definitions of fairness in terms of formal fairness “objects” (what we want to be fair to) and specifications (what being fair means) in the context of navigation planning. We use the same walkthrough examples to draw out challenges and design choices
50 that have to be thought through when deploying such methods. We discuss what the responsible development of fair planning methods requires in Section 5.2, a research agenda in Section 5.3 and the methodological contribution of the paper in Section 5.4. We conclude with a general summary in Section 6.

2. Fairness and responsible innovation

55 2.1. *Responsible innovation*

The growth of ethical guidelines for AI demonstrates a commitment to good practice across academia, industry and policy. However, it does not in itself guarantee ethical practice and does not necessarily specify how ethical conduct can be achieved. Winfield and Jirotko [4] argue for an agile and inclusive approach to ethical governance by drawing on the field of Responsible Innovation (RI). RI emerged from concerns surrounding the societal and ethical consequences of novel technologies [6] and has gained prominence in recent years as an EU and a UKRI¹ initiative that focuses on practices of “responsible development” in scientific research and ICT. The RI approach serves to explore and develop the means by which societal and ethical concerns can be identified and addressed throughout processes of research and innovation. Central to RI is to enable an inclusive, reflexive and accountable research and innovation process. This is for the most part achieved through the involvement of relevant stakeholders throughout the entirety of the research and innovation life cycle [7]. It emphasises the need to be sensitive to local, social and cultural contexts in the application of new innovations, to acknowledge the perspectives of relevant stakeholders and to recognise the importance of timing in the introduction of new measures that affect large groups of people.

A further aspect worth stressing is that RI is about anticipating issues, taking into account wider social, ethical and environmental issues and being able to create flexible and adaptive systems to deal with these unintended consequences—or what is known as anticipatory governance. Crucially, the RI approach is proactive and preventive rather than reactive, and is not intended to constrain. Instead it serves to help individuals and organisations to ensure the acceptability and societal desirability of research and innovation by influencing the trajectory of development early on in the innovation process. Thus, it can be

¹United Kingdom Research and Innovation

seen to shape a creative space in which researchers and innovators can generate insights informed by and aligned to societal and ethical concerns. This benefits the development of innovation and increases the likelihood of its acceptance in
85 society. RI principles are highly compatible with well-established practices in Participatory Design. The two may be combined [8] to include a broad range of stakeholder perspectives and encourage reflexive awareness amongst developers of their own role in the design process.

RI can offer insights into the practical application of robotic technologies in
90 order to understand how automated processes might be used in heedful ways that are ethically justified and do not compromise individual or group well-being. It avoids overly simplistic “one size fits all” solutions and adds an appreciation of context to abstracted discussions of ethics. The notions of responsibility and fairness are core aspects of the field, and can be seen to illuminate questions over
95 how an autonomous system might be designed to make “fair” decisions. Indeed, existing work conducted from an RI perspective [9, 10] has highlighted that fairness considerations are crucial to stakeholders when considering the application of autonomous systems. Similarly, questions of fairness, and controversies over lack of fairness [11, 12] have dominated public discussions of AI ethics.

100 Despite a wealth of academic literature on the subject, it is very difficult to formulate a working definition of what is actually constituted by “fairness”. As we will see in Section 2.4, there are multiple theories of fairness only within the philosophical literature. In relation to automated decision making, discussions of fairness in the literature typically make reference to some or all of the following
105 factors: moral responsibility of human actors; controlling the problems caused by automated processes; preserving the effectiveness of the technologies being used; and reducing undesirable outcomes. Rather than applying only a pre-set definition of fairness, we argue that it is possible to unpack what the concept means within a particular scenario. Drawing on the perspectives of RI we can
110 examine fairness in a way that is context specific and that focuses on how fairness might be pursued in a practical sense.

2.2. Inequality

Recent studies have shown that machine learning algorithms perform better for some people compared to others. For example, the popular investigative journalism article from ProPublica [12] revealed striking statistics regarding a recidivism prediction algorithm in use, called COMPAS [11, 13]. Recidivism prediction algorithms are algorithms that try to predict the likelihood that released criminals will re-offend. The article showed that black defendants were more likely to be wrongly predicted to re-offend, and white defendants were more likely to be wrongly predicted not to re-offend. Studies of facial analysis algorithms [14] have also shown the presence of large disparities in performance of commercial gender classification systems across gender and skin-tone. They reported up to 34% higher misclassifications on darker-skinned females compared to lighter-skinned males. They also found a large bias towards lighter-skinned subjects in related datasets. Similar disparities in performance were also shown to exist in facial analysis algorithms between healthy older adults and those with dementia [15]. Yet another example of measured disparities in algorithm performance is [16], which shows that state-of-the-art pedestrian detectors have considerably higher miss rates on children. The consequence for mobile robots such as autonomous vehicles (AVs) is that children could be more likely victims of accidents with AVs were they to be deployed with such algorithms.

Examples of social inequalities produced by seemingly fairness-unrelated decision-making are numerous and extend well past recent machine learning developments. In the field of “environmental justice”, for example, researchers have argued that the implementation of some transportation policies, such as a new metro system in San Francisco [2], which could supposedly improve mobility and access to jobs, can indirectly reinforce inequalities of opportunities—in particular deteriorating the access to transportation and the job market by low-income groups. Other work has shown that waste management sites are often concentrated on low-income, and high racial-minority-percentage locations [17]. Often such policies do no overtly target such populations, and inequalities of access or exposure to harm can happen because people are not uniformly dis-

tributed across space and are in fact usually distributed in ways that relate to economic, cultural and racial factors [2, 17]. Discrimination is also often embedded within housing markets and the organization of institutions [17], which
145 can implicitly influence decision-making and decision outcomes.

These particular discussions of spatially-organized inequalities also relate strongly to mobile robotics. The goal of autonomous vehicles is supposedly to improve access, quality and/or safety of transportation, but could come with
150 extra costs of reinforcing inequalities of access, pollution exposure, or others. Other service robots such as guide or shop-assistant robots may also provide differential benefits across hospitals or shopping areas. If spatially-organized inequalities can be argued to be unfair (which they can as we will see later), similar claims of (un)fairness could also be made about the way mobile robots
155 are programmed to navigate environments, because of inequalities they might produce. We will discuss this in more detail later in Section 3.2.

2.3. Indirect discrimination

The previous examples of inequality were measured along a set of categories, or personal characteristics that were deemed relevant to the task. In recidivism
160 the focus was on race, in facial analysis the focus was on race, gender, and age, and in environmental discrimination the focus was on race and income. In all the examples the inequality was (supposedly) not produced deliberately but resulted from correlations within a dataset or within the spatial distribution of people. They are, therefore, a form of indirect discrimination. Indirect discrimination
165 occurs when a policy or decision, while not explicitly targeting specific people or groups, has worse effects on people of a particular group [18].

The assumption behind all these examples, however, is that such categories are in fact morally relevant. There is the assumption that we know which features of individuals matter and which don't when evaluating fairness. These
170 features are usually called "protected characteristics" in the literature of political philosophy, and the motivation behind the choice of the categorization (gender, race, etc), at least in the "fair machine learning" literature is often

stated by the authors to be a legal one [19]. Indirect discrimination is legally protected in some countries through a specific set of characteristics—age, gender, religion, etc.—though the actual list and scope of the characteristics varies from country to country. A further claim made when choosing protected characteristics is that they should be those personal characteristics that people cannot control, such as to respect their autonomy [20].

2.4. Fairness claims and principles

Claims of “unfairness” of a certain measured inequality assume specific normative views of what a fair distribution is. One such view of distributive justice is egalitarianism, which can be defined as the view that some morally relevant factor such as health (or harm, or an important resource) should be distributed equally across individuals or groups [21]. Although the definition of the concrete distributive principle is itself disputed and several flavours of egalitarianism exist [21], the common assumption is that some kinds of inequality in the distribution of a good or harm are wrong. We will now briefly introduce a few distinct egalitarian distributive principles that relate to the discussions and definitions of fairness in the “algorithmic fairness” and “environmental justice” literatures, close in spirit to this article. We will relate these to robot navigation later in the text.

In what is sometimes called “telic” egalitarianism, the view is that all individuals or social groups should have equal quantities of the good or harm in question. Therefore, inequality itself, for example through pairwise differences, should be minimized. As described in [21], this could be written formally in a two-group case for example as maximizing $u_1 + u_2 - \alpha|u_1 - u_2|$, where u_i is the utility (or reward) of person i . In the extreme, it could also be interpreted as the need to enforce the constraint of an equal distribution, e.g. $u_1 = u_2$. Researchers in machine learning assume such a principle when they propose a loan-risk classifier for example, which predicts defaults equally across clients of any race (see experiments in [19] for an example). Basically, it enforces independence between a class label (e.g. gender) and a decision (e.g. to deny a loan).

The principle is often called “demographic parity” in machine learning and is enforced as a way to ensure a method “does not discriminate” [22]. This is also
205 the principle implied when, in the work of [23], the authors promote equal availability of taxis in all pick-up stations in an automated vehicle routing algorithm, though what is being equalized in that example are locations and not personal characteristics of the people involved. Often telic egalitarian arguments are also used in urban planning to claim that environmental harms such as exposure
210 to nuclear risk or waste facilities should be distributed equally across space or communities [17].

“Demographic parity” is criticized for failing to account for factors that people can control and that can morally justify inequality. For example, an algorithm for screening job applications should, according to these critics, only
215 promote equality among applicants that have similar experience to each other, e.g. number of years of education [24, 19]. Such views are in line with luck egalitarianism, a philosophical theory of distributive justice that says that inequality is only unfair when it relates to factors that arise by chance (called “brute luck” in the original theory of [25]), i.e. factors that are beyond personal
220 control. It is with such a normative assumption that, for example, [24] uses similarity metrics across individuals to make classification “fair” while treating similar individuals similarly.

In the previous principles the idea is to maximize average utility and avoid inequality. However, some philosophers argue that what matters is the utility
225 of the people that are worst-off. One such view was introduced by John Rawls in [26]. Rawlsian egalitarianism claims that inequalities are permissible insofar as they increase the wellbeing of the worst-off. This can be formally modelled as a maximin: maximize $\min(u_1, u_2)$, although other interpretations also exist [27]. This is the normative assumption made when ML researchers use such
230 claims that the performance of an algorithm “is only as good as the performance on the worst-performing group” [28]. Yet another fairness definition in the fair machine learning literature is explicitly inspired by Rawls theory and maximizes the utility of the worst-off group while constraining the inequality of regression

error across groups to be under a threshold [27]. One can imagine how such a
 235 principle could be important in autonomous vehicles if we wanted to increase
 access to the most isolated populations of a city—we will formalize this principle
 for robot navigation later in Section 4.3.

A similar principle is “sufficientarianism”. Although multiple versions of
 the principle exist [21], the common assumption is that fairness implies giving
 240 priority to people below a certain threshold of utility. This is closely related to
 discussions of the right to clean air [17], in environmental justice, or right to
 minimum service in public transportation. One particular version of sufficien-
 tarianism, from Skorupski [29], explicitly claims that fairness is achieved when
 utility is maximized at the same time as enforcing a bound on minimum utility
 245 applicable to all individuals.

Finally, prioritarianism is the view that “benefiting people matters more
 the worse off those people are” [30]. Formally, this can be written as maximize
 $g(u_1) + g(u_2)$ where g is a strictly concave function (e.g. $\sqrt{u_1} + \sqrt{u_2}$). This means
 that we maximize the sum of wellbeing over a population, but that wellbeing has
 250 diminishing utility (matters less and less) as it increases. This view is argued
 for in healthcare settings [31], where it can be said that what matters is not
 how worse-off someone is compared to others, but in comparison to what he or
 she could be (i.e. the degree to which they are in a bad condition) [30, 21].

2.5. Fairness and context

255 One thing that is clear from the discussions of fairness in the philosophical,
 environmental justice, and technical communities tackling fairness is that dif-
 ferent conceptions of fairness are incompatible with each other, and that the
 principle to be applied depends on the context or task at hand. Recent pa-
 pers also discuss how different fairness metrics are incompatible with each other
 260 [32]—only in very specific cases is it possible to have zero false positives and
 zero false negatives at the same time. Similar impossibility theorems have been
 proven in the social-choice (i.e. voting) theory literature as well, proving that
 no electoral system can simultaneously satisfy a set of three different fairness

criteria [33].

265 In addition, fairness can mean different things to different stakeholders. The
recidivism prediction case in particular generated discussions about whether
the algorithm was fair or not. On one hand the algorithm did not satisfy equal
false-positive or false-negative rates [12]. But on the other hand, the developers
of the algorithm defended the algorithm because it satisfied predictive parity:
270 true-positives and true-negatives were equal for white and black defendants [11].
They argued from the point of view of a decision-maker in the judicial system,
saying that error rates are of “no practical value” when predicting whether a
criminal will recidivate. However, from the point of view of the defendants, it
matters more that they are not wrongly classified as high risk [34].

275 Yet another finding is that, in some contexts, applying certain strict fairness
constraints can lead to the worse-off group being left at even worse utility levels,
just for the sake of balancing error rates [27].

And finally, which personal characteristics are morally relevant for the task
at hand also depends on the context. Being fair to the “age” of people might
280 make sense in some healthcare decisions but perhaps not in the context of hiring
or ad-targeting. Considering membership to the class “smoker” might make
sense in some healthcare decisions (even though this is also disputed) but not
in hiring or recidivism algorithms, etc. In Section 3.2 we discuss similar issues
of context in robot navigation.

285 2.6. *Fairness trade-offs*

Promoting fairness comes with its trade-offs. For example, enforcing a strict
fairness constraint such as “demographic parity” (i.e. independence between
a decision and group membership) while maximizing task-specific efficiency is
basically constrained optimization [35]. Clearly, achieving the same level of effi-
290 ciency in a constrained version of an optimization problem can only be done in
very specific circumstances. Some researchers in the fair machine learning com-
munity explicitly compare the efficiency-fairness trade-offs in different datasets
and fairness metrics by plotting Pareto curves [36]—showing the minimum fair-

ness constraint violation obtainable for each level of achieved efficiency. We will
295 in Section 4.4 use these to characterize a robot navigation problem.

One additional burden of enforcing fairness across groups, discussed by [35],
is the need to collect and use such group-membership data, and of applying
group-specific thresholds when making (hiring, loan, or other) decisions, which
could break core discrimination-related laws.

300 **3. Inequality and fairness in robot navigation**

3.1. Walkthrough rescue-robot example

We will now turn to discussing concepts of inequality and fairness in the
context of robot navigation. As a walkthrough example throughout the paper
we will draw on a hypothetical robot navigation planning problem, which we
305 describe next.

Imagine a robot that is deployed in the aftermath of a disaster in order to
find victims that need to be rescued or receive support. This could be a drone
searching for victims of an earthquake who need to be brought to a hospital, or
to receive medical or food supplies. The drone departs from a base station in
310 the city centre and needs to go back to the same station for re-charging batteries
and re-loading supplies after a maximum distance is covered. The drone does
several of these trips, although we focus on a single trip in isolation.

We will make the example numerical and grounded in data. Let us say this
happens in a specific city—Oxford, UK—and we use census data to guide the
315 robot towards high population-density areas to increase efficiency.

Figure 1 shows Oxford’s spatial distribution of population density, age, eth-
nicity and gender. We will focus on these three variables as they are considered
“protected characteristics” in many contexts, as described in Section 2.3. As
the figure shows, population density is not uniformly distributed in space. Fur-
320 thermore, it has spatial biases related to age, ethnicity, gender, etc., as we have
discussed in Section 2.4 when introducing the literature of environmental justice.
Specifically, the city centre of Oxford is densely inhabited mainly by students

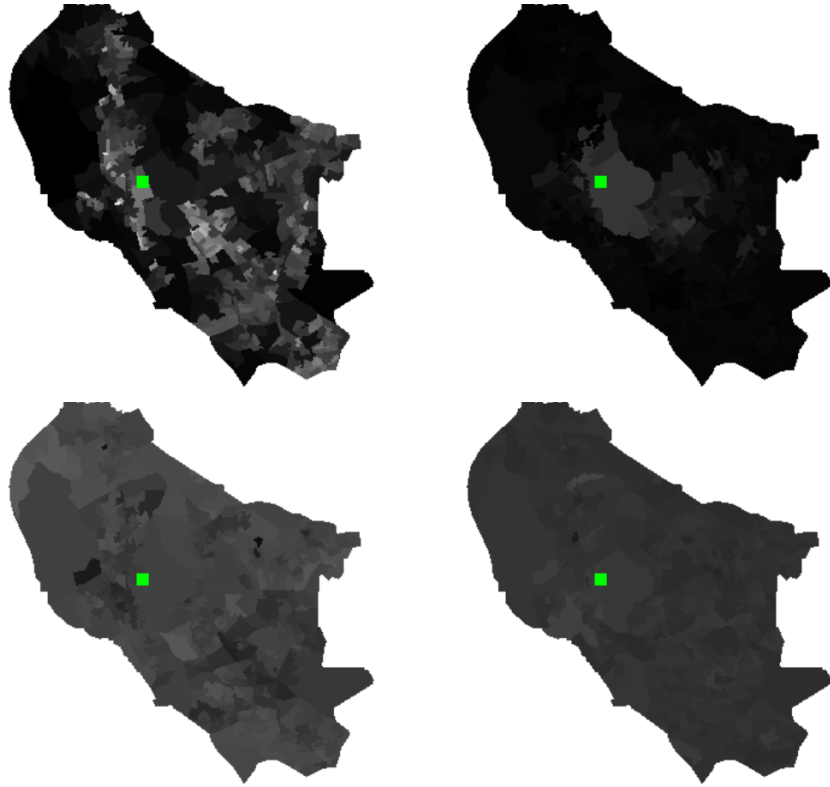


Figure 1: Data used for robot navigation simulations. Population density (top left), percentage of people aged 20-24 (top right), percentage of people of ethnicity “white English” (bottom left), percentage of people of gender “male” (bottom right). Higher values are brighter. The robot’s home base is marked with a green square.

in their 20s, mainly of white English ethnicity. Like other cities [2], Oxford has neighbourhoods of higher concentration of minority ethnicities and of older populations than the city centre.

The consequence for our rescue drone example is the following: planned paths will have skewed distributions of these personal characteristics. If for example the drone thoroughly explores the area immediately around the base station it will find many people because of population density, however most of whom are young and healthy. These could arguably survive longer without drone help, compared to older populations further away from the centre. A

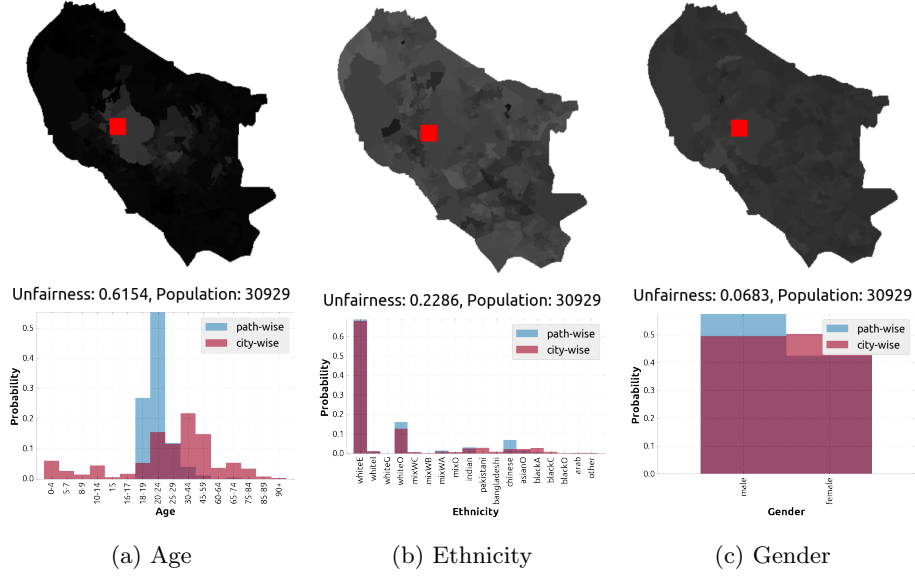


Figure 2: Distribution of age, ethnicity and gender over the whole city and a robot path. The path is a naive thorough search around the home-base.

decision- or policy-maker could have similar or other claims with regards to various personal characteristics.

Let us suppose that the navigation planner is such that it thoroughly explores the region around the base-station because it is highly populated. Figure 2 shows this path on top of the city map, as well as the path-wise and city-wise personal characteristic distributions of age, ethnicity and gender. Figure 2 shows that, as qualitatively seen on the maps, the distribution of age in the centre (along the robot’s path) is highly biased towards that of undergraduate students, while the city-wide distribution is considerably more uniform. The figure also shows that both the centre and the city as a whole are highly biased towards a “white English” ethnicity. The centre has an overrepresented “white other” and “Chinese” population compared to the rest of the city.

3.2. Issues raised by this example

345 3.2.1. Indirect discrimination within robot navigation

The previous example shows that robot navigation paths can be biased in favour or disfavour of different people, as paths inherit spatial distribution biases. In particular, the probability of being found by the rescue robot was strongly correlated with age and ethnicity (i.e. no demographic parity).

350 In disaster response, such a robot could continue or even reinforce common criticisms in disaster response missions: that policies for selecting disaster response locations usually benefit particular groups of people [37]. Avoiding such discrimination explicitly through algorithms could be a way not only to promote distributive justice, but also to enforce a certain degree of political or commercial neutrality in disaster response (i.e. to make sure that disaster response
355 agencies using robots do not favour any particular group of people).

Indirect discrimination will happen in many other robot scenarios as well. Consider one timely example use case: an autonomous vehicle (AV). If the AV applies different prices-per-distance depending on expected traffic conditions, this could indirectly penalise people working in certain areas of a city
360 and further reinforce inequalities of access to jobs. If the AV applies different prices-per-distance depending on predicted route risk, such as the probability of vandalism or theft, then this could penalise people living in high crime-rate areas. Such schemes could introduce or reinforce social inequalities, similarly to
365 previous cases of unintentionally discriminatory infrastructure and urban planning practices [2].

3.2.2. Inequality can be unfair

While one of the goals of a disaster response robot is to find/rescue/treat as many people as possible, notions of priority also exist in disaster response ethics
370 [37]. One accepted principle is to attend to the people most-at-risk first [38, 37]. This is related to a prioritarian view of distributive justice, that a decision is fair when it optimizes total utility but some people’s utility is more important than others (i.e. people with the lowest current “utility”, such as those with low

health-state, those living in low-quality accommodation susceptible of collapse,
375 those that have lower chances of survival such as children and older adults,
etc.). This means that, in our particular example, the fact that there is indirect
discrimination of age, with a bias towards the younger population, is unfair
according to disaster response ethics.

So while our robot example is doing part of a disaster response team’s job—
380 that of finding as many people as possible—it is not respecting the context’s
notion of distributive fairness, the notion of who most needs to be found first
in the context of disaster response. Another way to see this is that this specific
algorithm is taking resources away from those that need it most.

This example also raises the issue of identifying in which personal charac-
385 teristics indirect discrimination is unfair. What about the case where a rescue
robot finds many people, primarily those at highest risk, but at the same time
is biased towards the white population of the city—since it is slightly closer
to the base station than the neighbourhoods of high minority-concentration?
Should this give rise to concern? According to defendants of affirmative ac-
390 tion it should, since it reinforces social inequalities that have been ingrained in
society (through urban policy, housing markets, etc.) for generations.

It then becomes important to provide stakeholders with tools to help identify
these possible inequality outcomes of robot deployment so that better design de-
cisions can be made before robot deployment. One such tool is the methodology
395 that we use in this paper—of simulating the robot’s deployment and predict-
ing possible inequalities that can give rise to concerns of distributive fairness.
The example further raises the issue, however, of how one will reach the “final”
choice of fairness principles and protected characteristics to use in an algorithm.
This requires involvement of all stakeholders in the decision process, as well as
400 other responsible innovation methods as we discuss in Section 5.3.

3.2.3. *Fairness depends on context*

As we have seen, in the rescue robot case, fair navigation planning would
require respecting fairness principles of rescue teams such as most-at-risk-first,

but also perhaps avoiding indirect discrimination. Just as in the discussions of
405 fairness in machine learning (Section 2.5), the definition of fairness is going to
depend on contextual factors such as the task and stakeholders. While rescue
robots should potentially focus on finding people that are most-at-risk, fairness
would probably mean something else in other use cases, for example:

1. Autonomous vehicles (e.g. taxis) with location-based insurance or trip
410 pricing. Because of the close link to public transport, here fairness could
be more closely related to sufficientarian “minimum service” policies in
transportation (Section 2.4). So fair navigation planning would in this
sense be to provide guarantees that paths taken by AVs provide fair levels
of waiting-times across all areas of a city. In addition to that the personal
415 characteristics to protect and equalize could be related to other factors
different from the rescue case, such as income level or crime-rate. One
could argue that pricing should not be correlated with area crime-rate
itself, even if it correlated with risk estimates for insurance purposes, as
this would lead to the double punishment of users—of having to pay extra
420 on top of already having to go through the exposure to high crime risk.
2. Hospital mobile service robot. Imagine this robot visits different areas
of a hospital while guiding patients and visitors to their destinations, or
executing tasks on their demand. Depending on how much time a robot
spends in each of a hospital’s waiting rooms, corridors and patient rooms,
425 people will have more or less chance of using the robot’s services. In
this context, hospitals most often have clear criteria of patient priority
based on their condition [39], and mobile robot services in a hospital could
reflect these as well. In this case, fairness specifications could involve
guaranteeing relative degrees of service, for example: guaranteeing that
430 robot paths service M times more patients of priority 1 than those of
priority 2. Alternatively, that male and female patients are given the
same amount of attention.

3.2.4. *Robots must face dilemmas that humans already face*

Thinking about the fairness issues related to disaster response locations is
not something new about rescue robots, but an inherent concern of disaster
response itself. For example, discussions and claims of unfairness are raised re-
garding disaster response hospital locations when they favour people of specific
backgrounds, sometimes politically or commercially favourable to the country
backing the response [37]. Principles such as most-at-risk first [38, 37] are pro-
posed within the disaster bio-ethics community to solve a fairness problem in-
herent to the task and context. Thus, the rescue robot example also shows that
when robots are used to solve problems currently solved by humans, they must
face similar dilemmas currently faced by human decision-makers. In the hospi-
tal robot example as well, hospital workers already face ethical questions about
how they spend their time in the physical environment of a hospital. The ques-
tion is then how to operationalize similar fairness sensitivity when automating
a service with a mobile robot.

Of course, new fairness issues may arise with the introduction of automa-
tion. For example, in the hypothetical example of an autonomous vehicle with
location-based insurance or trip pricing, the indirect discrimination issue is
partly new in the way it is personalized to a person’s travel needs or routines.
However, concerns of fairness are already an important part of the discussions
taking place regarding car insurance, and insurance providers are required to
meet certain (loose) criteria of fairness in their policies, such as avoiding direct
discrimination based on certain protected characteristics [40].

4. Designing fair robot navigation systems

4.1. *The objects of fairness in navigation: qualitative definitions*

In this section we try to list the kinds of things we might want to be fair
to in robot navigation. We distinguish between different objects of fairness
considerations in navigation:

4.1.1. Locations

Being fair to locations (e.g. hospital rooms) means that we care about which locations will be visited by a robot: in particular, how often they will be visited, either in absolute value or in comparison to other locations. We refer to location-related objects/variables technically as “state visits”, since the term “state” is frequently used to formally describe a location of the robot in motion planning literature. The work on formal verification and formal methods for robotics has traditionally focused on this kind of variable [41, 42, 43], particularly in logical events indicating whether a region is eventually visited, or visited infinitely often, by a plan.

4.1.2. Protected characteristics

The discussions reviewed in Section 2 reveal that there is often a need to be fair to personal characteristics (e.g. age or health) of the people affected by algorithms. And as we saw in Section 3 with the rescue robot example, inequalities across protected characteristics will also be part of the outcomes of mobile robot deployment. The other objects we consider here are, therefore, morally relevant protected characteristics. Being fair to protected characteristics, in the context of navigation planning, means that we care about the distribution of protected characteristics of the people found/interacted-with along the robot’s path. We refer to these technically as “state features”. In each location lie a certain (deterministic or stochastic) number of people, and the distribution of their protected characteristics will be part of the “features” of that state.

4.2. The objects of fairness in navigation: formal definitions

Let us assume our robot’s state-space is \mathcal{S} and for simplicity a path ζ is a discrete sequence of N states $\zeta = s_1, \dots, s_N$, where $s_i \in \mathcal{S}, \forall i=1, \dots, N$. We consider the state-space of the robot is divided into mutually-exclusive R regions $\pi_i \subset \mathcal{S}, \forall i=1, \dots, R$. A state-feature is a random variable $A \in \{A_1, \dots, A_C\}$ (e.g. for person gender, $C = 2$, $A_1 = \text{male}$, $A_2 = \text{female}$).

Formally, the objects of fairness we have just introduced can be written as:

- 490 • State visit counts: $c(\pi_k) = \sum_{i=1}^N \mathbf{1}_{[s_i \in \pi_k]}$, where $\mathbf{1}$ is an indicator function equal to 1 when the subscript condition is true and 0 otherwise.
- Eventual visits (a location is eventually visited by a plan, at least once): $\Diamond \pi_i$, where π_i here represents the logical event of a robot being present at region π_i , as used in Linear Temporal Logic [44].
- 495 • Infinite visits (a location is visited infinitely often by a plan): $\Box \Diamond \pi_i$.
- Distribution of A over the path: p_A^ζ , where $\sum_{a \in \{A_1, \dots, A_C\}} p_A^\zeta(a) = 1$. To use a notation closer to the machine learning literature, we are interested in the distribution $\mathbb{P}(A = A_i | Y = 1)$, where $Y = 1$ is the event that a person is found along the path and $A = A_i$ the event that a person's protected characteristic is A_i . In a two-class gender example, if the user has access to data of the expected number of male E_m and female people E_f present at each location in a map, then:

$$p_A^\zeta(A_m) = \frac{\sum_{i=1}^N E_m(s_i)}{\sum_{i=1}^N E_m(s_i) + E_f(s_i)}, \quad (1)$$

$$p_A^\zeta(A_f) = \frac{\sum_{i=1}^N E_f(s_i)}{\sum_{i=1}^N E_m(s_i) + E_f(s_i)}.$$

4.3. Specifications of fairness in navigation

Fairness principles can be implemented as optimization problems with costs and constraints on the variables just described.

The form of fairness currently studied in the planning literature is LTL fairness. These are logical specifications on state visit events, such as “if region 1 is visited then region 2 must also eventually be visited”, or “if 1 is visited infinitely often then 2 must also be”. The first case could be specified in Linear Temporal Logic (LTL) by $\Box(\pi_1 \Rightarrow \Diamond \pi_2)$ which is usually called “liveness” [44, 41]. The second specification is called “strong LTL fairness” and is represented by $\Box \Diamond \pi_1 \Rightarrow \Box \Diamond \pi_2$. These definitions can be seen as a person-agnostic egalitarianism equalizing visits over sub-sets of locations. However, as we have seen there are other notions of fairness that may apply to navigation problems.

515 Here we collect from the discussions in Section 2.4 (and translate into the robot navigation domain) other fairness specifications relevant to navigation:

- Demographic parity. In the context of navigation this egalitarian principle implies that the event of a person being found along a robot’s path is independent from group membership (i.e. a protected characteristic A).
520 **Formal:** A constraint $p_A^\zeta = p_A$, which enforces the path-wise distribution of features to be equal to the population-wise distribution. Equivalently and closer to the machine learning definition $\mathbb{P}(Y = 1|A = A_i) = \mathbb{P}(Y = 1|A = A_j)\forall_{i,j}$.
- Rawlsian egalitarian fairness: maximizing utility of the worst-off, i.e. maximizing the visit counts of the least-visited region, or the probability of being found for the least-likely group. **Formal:** For state-visits, the goal of the planner is to maximize $\min_{i=1,\dots,R} c(\pi_i)$ **Formal:** For state-features, the goal of the planner is to maximize $\min_{a \in \{A_1, \dots, A_C\}} p_A^\zeta(a)$.
525
- Sufficentarian fairness (1): Requiring a minimum number of visit counts at specific regions (e.g. “minimum service” constraints in a hospital).
530 **Formal:** A bound-constraint of the type $c(\pi_k) > \phi$ where ϕ is a user-specified threshold.
- Sufficentarian fairness (2): Requiring a lower-bound on the relative number of visit counts (e.g. minimum fraction of attention given to a specific room) or the probability of each group being found by the robot (e.g. all age groups must have at least probability 0.1 of being found). **Formal:** For state-visits, a constraint $\frac{c(\pi_k)}{\sum_{j \in \mathcal{J}} c(\pi_j)} > \phi$, where J is a user-specified ratio and \mathcal{J} a user-specified set of regions **Formal:** For state-features, a bound constraint $p_A^\zeta(a) > \phi$ on all classes $a \in \{A_1, \dots, A_C\}$.
535
- Prioritarian fairness: Maximizing the priority-adjusted utility of visiting locations or groups. **Formal:** For state-visits, the goal of the planner is to maximize $\sum_{i=1,\dots,R} g(c(\pi_i))$ where $g(\cdot)$ is a strictly concave function such as $\sqrt{\cdot}$. **Formal:** For state-features, the goal of the planner is to maximize
540

545 $\sum_{a \in \{A_1, \dots, A_C\}} g(p_A^\zeta(a))$ where $g(\cdot)$ is a strictly concave function such as $\sqrt{\cdot}$.

• Affirmative action: Enforcing a desired distribution of states-visits (e.g. that room of priority 1 is visited M times more than room of priority 2) or state-features (e.g. ratio of younger and older people found along the planned path to be as close as possible to 20/80). These are closely related
 550 to affirmative action policies in university admissions and public offices, hence the name we choose for them. However, note that this definition is general enough to encompass demographic parity—which is equivalent to a preference towards the distribution of the whole population. **Formal:** For state-visits, a constraint $\frac{c(\pi_k)}{\sum_{j \in \mathcal{J}} c(\pi_j)} = \phi$, where ϕ is a user-specified ratio and \mathcal{J} a user-specified set of regions. **Formal:** For state-features,
 555 a constraint on the distance between the path-wise distribution p_A^ζ and a desired distribution Q , e.g. $D_{KL}(p_A^\zeta || Q) = 0$.

Importantly, note that specifications based on state-features (i.e. distributions of protected characteristics) are non-Markovian, that is, they lead to a
 560 problem without optimal sub-structure. To see that this is the case, notice that the reward of moving from a state s_i to one of its neighbours will depend on how the robot reached s_i (what is the distribution of the features of interest at the moment). Optimal sub-structure is a requirement for typical state-of-the-art dynamic-programming based methods such as A* and value-iteration-
 565 based approaches to planning. This fact introduces a new need to research non-Markovian planning methods, and makes distribution-fairness definitions the most technically interesting.

To deal with non-Markovian cost functions in dynamic-programming based methods, some techniques have been proposed. Similarly to what is done in
 570 [45] for time variables, distribution-related data itself could be added as part of the state-space (e.g. the robot state is now location X , total number of people visited Y , feature ratios Z), though such methods typically do not scale well with state-space dimension. Another more straightforward approach is to

build a proxy planning problem with cumulative costs that tries to achieve a
575 similar objective. In Section 4.4 we see how this method fares against solving
the original (not approximated) problem.

In the experimental part of this paper we will focus on fairness to protected
characteristics, or state-features, for two reasons. One is the technical difficulty
it poses to existing methods. The other is its relevance to topics of “indirect dis-
580 crimination” on protected characteristics. As we have discussed in Section 2.3,
such considerations are both the object of legal regulations [46] and the topic of
analysis and active debate in the context of algorithmic bias and discrimination
[13, 19].

4.4. *Developing a fair navigation planner*

585 Now that we are equipped with a framework of fairness variables and speci-
fications within robot navigation, we are ready to analyse and discuss its applica-
tion to existing techniques for navigation planning. With a particular example
in hand, such as our rescue robot use case, we can begin asking important
questions such as: “what is the trade-off between fairness and efficiency in our
590 scenario?”, “is it possible to achieve fairness strictly?”, “is our fairness speci-
fication counter-productive?”, “can traditional navigation planning techniques
be applied to our problem?”. We will now try to answer such questions for our
particular rescue example and use it to generalize. For the sake of guiding the
discussion we will assume that, at a particular point in time in the stakeholder-
595 involved design process, we wish the probability of a person being found by a
robot to be independent of the person’s age, i.e. demographic parity. This could
be seen as an attempt to compensate for the large spatial segregation of the city
in terms of age (i.e. very young or gentrified neighbourhoods).

4.4.1. *Fairness may be infeasible, requires trade-offs*

600 To gather insights of the feasibility and trade-offs of enforcing demographic
parity in this hypothetical scenario, we estimate the Pareto-front of the two
objectives: fairness (distance to perfect demographic parity) and efficiency

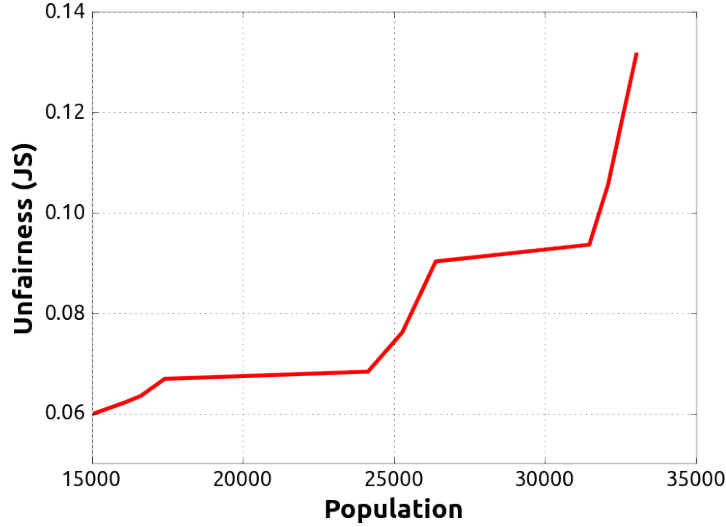


Figure 3: Pareto-curves showing the trade-off between the total population found along the robot’s path and the distance to the desired distribution of age.

(number of people found). For this purpose we use a state-of-the-art Pareto-estimation method adapted to navigation planning, which we describe in detail
605 in the Appendix (Section 7.2).

Figure 3 shows this Pareto-front. Each point along the curve is a different path of different fairness and efficiency. The graph shows that to decrease unfairness, which in this case means the distance to perfect demographic parity, it is necessary to reduce the total population found. In our example, this is
610 because older people also live more scattered and further away from the centre than the younger population. It also shows that it was impossible for the method to find a path of strict fairness, i.e. where demographic parity is satisfied exactly, since the curve does not reach zero. This is understandable because, firstly, extreme luck should exist for a subsample of the city to exhibit exactly the
615 same statistics as the city as a whole. Secondly, Pareto-front estimation methods such as the one used here are not guaranteed to find global minima since that would require exhaustively searching all possible paths within the map, which is unfeasible for the size of our problem.

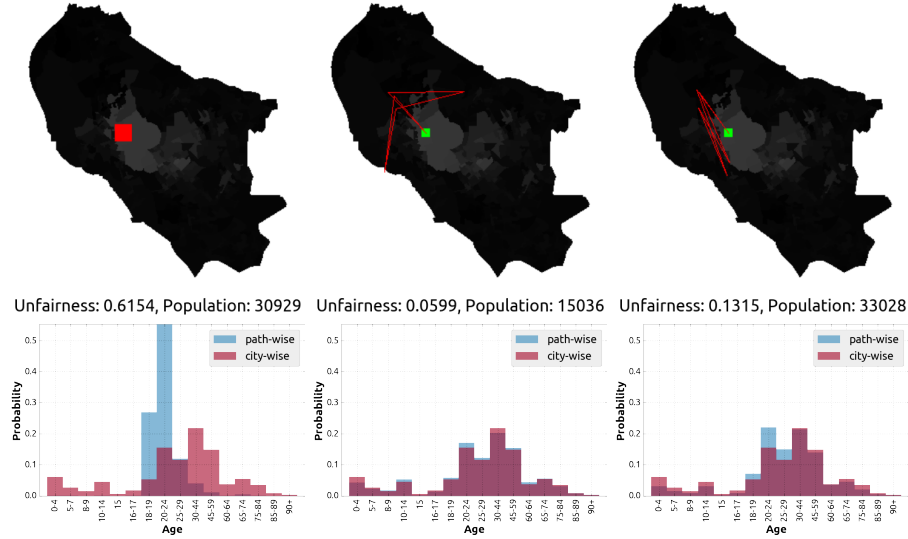


Figure 4: A thorough search path around the home-base (left) and the two extreme Pareto-optimal paths (middle and right).

Figure 4 shows the paths corresponding to the two extremes of the Pareto-front (i.e. lowest and highest unfairness). The method obtains paths with lower unfairness for the same amount of population found, when compared to a thorough planner focused on the city centre. The difference is considerable: more than 50% of the people found by the thorough planner are aged 20-24, while the city-wide percentage is around 15%. For our planner this number is between 17 and 22%. The histograms in Figure 4 show that both extremes of the Pareto-front capture a population which is close to the city-wide distribution (i.e. close to independence). The higher-unfairness extreme can actually find a slightly higher population than the thorough planner, at considerably lower unfairness. To achieve this, the planner's paths move through highly populated areas of both younger and older populations, scattering away from the base position more than the thorough planner.

In this situation, a stakeholder such as a decision-maker, emergency responder or policy maker could use the Pareto-fronts themselves to make a more informed decision about efficiency and fairness of the response, in a way that

635 reflects the priorities and values at play in the specific situation. The decision-maker could select one of the solutions within the Pareto and be comfortable that they represent approximately optimal trade-offs of the objective (though not globally optimal as already discussed). The choice of solution along the Pareto-front could also be made according to the resources available to the base station at each moment and expected needs of the different kinds of population.
640

4.4.2. *Current planning methods provide few guarantees*

The method we used to generate these paths is not a traditional one in robot navigation: because the focus in robotics is usually not on trade-off estimation, and because faster methods with optimality guarantees are preferred. However, such traditional methods (e.g. A* with an admissible heuristic) require cost
645 functions to be Markovian, i.e. the cost over a path to be equal to the sum of per-state costs, and state-feature definitions are not decomposable in such a way. So even though methods with optimality guarantees could also be used here, they would be optimizing a proxy function of fairness, i.e. would guarantee that they would find the path of maximum fairness but where fairness is not
650 the actual definition we care about.

To analyse the results of applying such a traditional planner with a Markovian approximation of fairness, we conducted a new experiment. We approximated the fairness over a path by the fairness within each cell summed over the whole path:
655

$$\hat{f}_{\text{unfairness}}(\zeta) := \sum_{(s) \in \zeta} f_{\text{unfairness}}(s), \quad (2)$$

This way, our problem becomes solvable with traditional state-of-the-art search- and sampling-based planning methods [47, 48] with guarantees. The danger here is that promoting each cell’s feature distribution to be as close as possible to the city’s distribution could be too restrictive. It could lead to avoiding minority regions because they are too different from the city’s average. Figure 5 shows the original Pareto-front and age distribution, and those obtained by optimizing the proxy problem. All graphs and unfairness metrics shown are those of path-wise
660

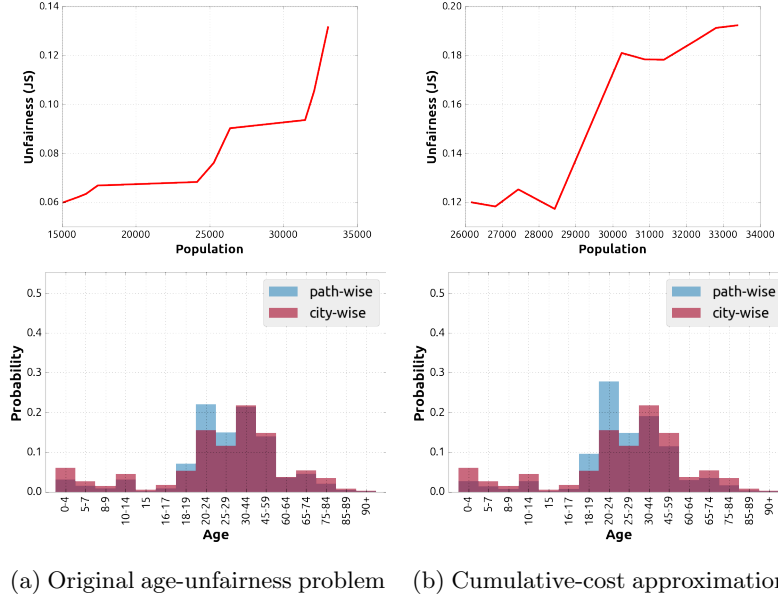


Figure 5: Comparing the original age-unfairness problem (a) with the result of solving a proxy problem which minimizes approximate cumulative-costs of per-cell-unfairness (b). The histograms in the bottom correspond to the highest-population extreme of the Pareto front.

and city-wise distribution distances, the only difference is in the cost function that was used to arrive at these results. Notably, the cumulative-cost approximation leads to twice higher unfairness (i.e. distance to perfect demographic parity) related to a bias towards the younger population. The high-population extreme of the front finds a similar number of people (around 33000), though the low extreme is lower in the case of the original cost function. These differences could be more or less pronounced depending on the problem, and it could turn out to be reasonable for a decision- or policy-maker to accept these degrees of unfairness. In that case, other state-of-the-art methods using cumulative-cost approximations could be used, perhaps with some formal or optimality guarantees. Still, there would be a danger then, that problems where this approximation turns out problematic would go unnoticed. Additionally, the use of traditional methods with formal or optimality guarantees on surrogate functions could lead to undue trust in the fairness of the system, perhaps specially

because of user knowledge of those guarantees. Therefore, we argue that the use of any fairness-aware planner should always go together with visualizations of fairness and other tools that promote responsibility.

680 To summarize, current navigation planning methods either provide optimality guarantees for the wrong metric, or provide no guarantees on the desired fairness metric. Optimality guarantees are an important feature that allows technology users to be sure that the technology does what it is meant to—optimize fairness and efficiency as much as possible. Not having these guarantees means that the users do not know how much efficiency and fairness they
685 are potentially losing in each use.

4.4.3. *Fairness specifications can be counter-productive*

To evaluate the possibility of counter-productive fairness constraints, consider an egalitarian view that all classes of a protected characteristic should
690 have equal probability of being present in the robot’s path. Figure 6 shows the result of optimizing for this uniform-distribution affirmative-action on age and ethnicity. The minimum achievable distance to strict “fairness” is very high in this case: at least 0.46 on age and 0.7 on ethnicity. While the majority classes were considerably lowered, the outcome is still very far from the preference. In
695 other words, in this case the city’s inequality is just too high for a uniform distribution of characteristics to be achieved. This means that to achieve a certain degree of “fairness” may require an extreme decrease of efficiency if the metric is not well chosen. Now consider a similar policy applied on gender classes. Figure 7 compares the largest-efficiency result of promoting demographic parity
700 on gender, with the largest-efficiency result of promoting a strict-equality affirmative action (i.e. a 50-50% ratio). The figure shows that strict-equality leads to lower utility for *both* male and female classes. In other words, all classes were made worse-off just in order to find higher equality solutions.

Again, the use of fairness-related visualizations such as distributions and
705 trade-offs is important to inform decision-makers about the impact and effectiveness of the technical choice. Furthermore, this example shows the advantage

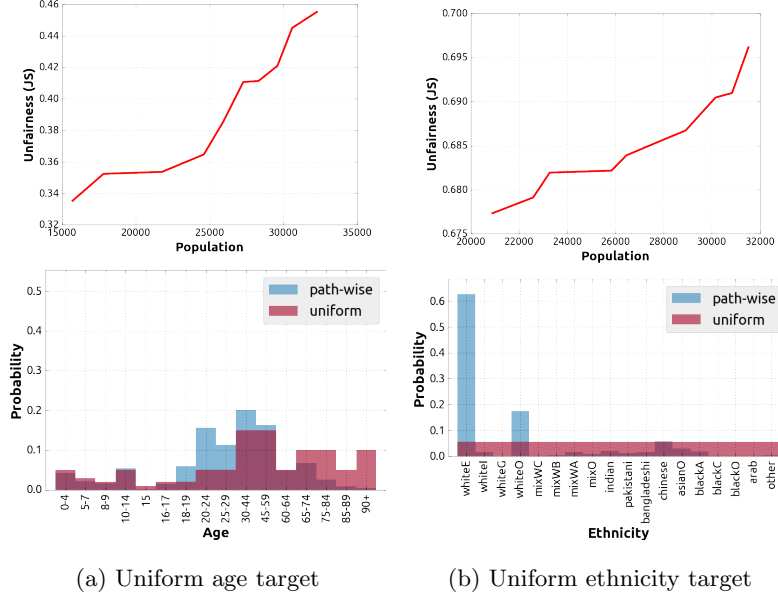


Figure 6: The hopelessness of achieving some target distributions, in this case “fair” uniform distributions. Note that in (a) the age ranges for each bar (labels on histogram x-axis) are not of constant size, which is why the distribution does not look flat.

of our methodology, of simulating system deployments implementing different fairness definitions, in order to better evaluate and predict their results. Finally, the example suggests that the process of specification—of selecting a definition of fairness—could be the product of an iterative design and validation process.

4.4.4. Intuitive understanding

The particular metric we have used for “unfairness” in Pareto-curves—distance to the fair distribution—also raises questions about which metrics are most intuitive for stakeholders to evaluate the degree to which a distributive principle is satisfied. In this example we used Jensen-Shannon distance between distributions, but such a metric could arguably be considered unintuitive. Other options could be the maximum violation of the fairness constraint as used in [36]. In our example this could amount to evaluating the maximum pairwise difference of the probability of being found conditioned on group membership

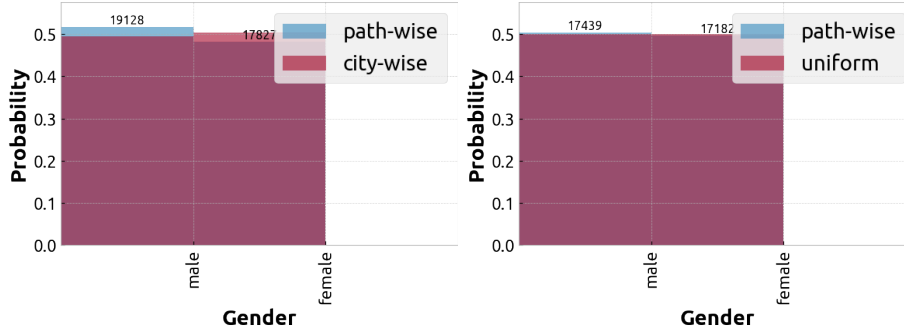


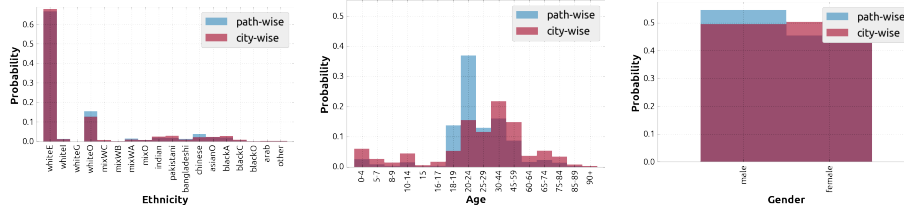
Figure 7: The implementation of certain fairness definitions can be counter-productive, i.e. achieve more equality at the cost of less utility for every group. Left: promoting demographic parity on gender. Right: promoting affirmative action (i.e. a 50-50% ratio). The numbers over each bar represent the number of people found along the robot’s path.

(i.e. $|\mathbb{P}(Y = 1|A = A_i) - \mathbb{P}(Y = 1|A = A_j)|$). The best metric would probably depend on the fairness specification itself and should be the subject of research undertaken in collaboration with social scientists and Human-Computer Interaction researchers. We discuss in greater detail in Section 5.3.

4.4.5. Design is an iterative process

A further issue that the rescue robot example raises is that there are possibly multiple personal characteristics that a user or decision-maker cares about and would like to pay respect to. These may not be obvious from the onset of robot deployment. For example, a disaster response team might program a fair navigation planner to respect a certain health and age-related feature, only to later find out they have a bias towards high-income neighbourhoods that they would like to avoid. Alternatively, optimizing for fairness in one characteristic may introduce new biases in the paths that are again morally relevant.

Figure 8 illustrates part of these ideas. We show the result of optimizing for ethnicity-fairness, both in terms of obtained ethnicity distributions and the resulting distributions of the (unconsidered) characteristics of age and gender. The figure shows that this path is still biased in terms of age and gender. The visualizations make this fact clear, however, and could trigger a stakeholders to



(a) Distribution of ethnicity over path (optimized). (b) Resulting distribution of age over path. (c) Resulting distribution of gender over path.

Figure 8: Conflicts between fairness objectives. When minimizing ethnicity-unfairness, the resulting distributions of other characteristics may (still) be undesired.

re-consider what the objectives of the planner should be.

Part of the specification process will hence be in the discovery of what values
 740 matter for (and are affected by) the application at hand. This process can
 only succeed if there exist value-sensitive tools [49] that identify or support the
 identification of these issues. Our methodology in this paper was to develop a
 resource for developers of applicable formal definitions that can be simulated to
 guide discussions and anticipate issues *before* deployment, at the early stages of
 745 development.

5. Discussion

Thus far we have given an overview of concepts of responsible innovation,
 inequality, discrimination and fairness in different literature (Section 2), and
 related those to robot navigation through an example use case (Section 3). We
 750 have claimed that indirect discrimination is likely to arise in navigation, and
 that robots will have to deal with (or allow humans to have control over) similar
 fairness-related questions that currently arise in non-automated instances of the
 navigation problems. We demonstrated the use of a fairness-aware algorithm
 on the rescue use case, and raised some issues regarding the development and
 755 deployment of such algorithms.

5.1. Related work in planning

The ethical and social focus of this paper in planning and navigation is in line with other recent work in socially-aware motion planning. For example, [50] selects robot motion waypoints in a way that is fair to different human tele-operators, [51] considers the violation of personal space of closeby pedestrians, and [52] considers user privacy within algorithms for vehicle routing. While in this paper we focused on optimizing fairness and task efficiency objectives, motion planning usually deals with optimizing objectives such as energy, time, safety, etc. Multi-objective approaches to motion planning already exist and include the use of weighted-sum cost aggregations [53, 54] or obtaining Pareto-curves that optimally trade-off the different objectives [55, 56, 57, 45]. Pareto-front estimation has been applied to both search [55] and sampling-based [56] motion planning, as well as formal methods for planning in Markov Decision Processes [45] and reinforcement learning [57]. Other relevant work is that of [58], which applies different distributive principles to multi-objective planning such as elitist and Rawlsian egalitarian principles. In this paper we also use Pareto-optimization, both to obtain optimal trade-offs and to allow for an intuitive understanding of those trade-offs to stakeholders through visual inspection of the Pareto-fronts.

5.2. Developing fair navigation planners responsibly

As our walkthrough example has shown, building “fair navigation planning” algorithms in a responsible way certainly requires more than optimizing a fairness metric. Based on the previous discussions and example-based observations we now sketch what we believe the ethical development and deployment of fair navigation algorithms requires. We use a set of principles common across the recently proposed ethical AI frameworks [5] to guide the ethical requirement themes: transparency, privacy, autonomy (in the philosophical sense of humans being in control) and fairness.

5.2.1. Transparency: Providing an understanding of inequality and trade-offs

785 Planners should provide an intuitive understanding, through appropriate visualizations, statistics, or metrics, of the fairness characteristics of navigation plans. From a Value-Sensitive-Design perspective [49], it is important that our tools provide us with information that lets us understand and control the impact of the technology in terms of the values of interest (fairness and others). In terms
790 of algorithm and interface design we should not just be turning a fairness feature on and off, but providing data and options so that a stakeholder can correctly understand and act on the fairness dimension of the problem. Planners should be equipped with data and tools for the analysis of fairness across multiple variables and specifications, in order to allow impact and fairness-related issues
795 to be spotted by stakeholders.

Legibility or intuitiveness of the visualizations and the metrics is important as well. For example, in this paper we used Jenson-Shannon distance of distributions as a way to measure the degree to which the fairness definition was satisfied. Alternative metrics such as the maximum violation of the constraint
800 (e.g. maximum difference of the probability of being found across group membership) have been used in the literature [36], and the choice of such metrics could have an important role in transparency and understanding of trade-offs and predicted outcomes of system deployment.

5.2.2. Human autonomy: Providing control over fairness and efficiency trade-offs

805 In order to allow humans to responsibly use such tools, they need to be able to understand and control the trade-offs between fairness and other task objectives. This will require the use of visualization and human-in-the-loop design features. For example, Pareto-curves of the different objectives can serve as
810 interesting visualizations and decision-aids. Some Pareto-estimation techniques already exist for both generic optimization problems [59, 60] and planning problems [45], however their computation is still slow, sub-optimal or challenging with the increase of the number of objectives. Research to improve this reality,

and research on alternative visualization and decision-making tools is needed
815 to make fair navigation planning techniques useful and reliable. Additionally,
users should be able to interface with the planning methods in order to include
considerations of relevant expert knowledge, such as adding intermediate goals,
biasing paths to certain solutions, or adjusting estimates of risk and utility (e.g.
building damage or population density in the rescue case).

820 5.2.3. *Privacy: data collection, security and inference*

Promoting state-feature fairness in navigation, such as demographic parity
or other, requires data on the distributions of these features themselves. This
comes at the cost of having to gather such data, but also of potential privacy
issues within the collection, analysis or security breaches of the data. Data
825 leaks could come not only from security breaches but also correlations within
observed robot behaviour. Ideally, paths taken by a mobile robot should also
not reveal information about the personal characteristics of the people in the
city or respective deployment location. This should be true even if the formal
definition of fairness used by the robot is known by the public for the sake of
830 transparency. This concern for privacy in motion planning has been recently
discussed in technical papers such as [52], and fair planners will likely require
to go hand-in-hand with such privacy-assuring methods.

5.2.4. *Verifiability: Providing formal and optimality guarantees*

Users may need proof that the system will work as intended, both in terms
835 of achieving the fairness specification that is asked, and of being as efficient
as possible. Formal methods and probabilistic model checking are useful tools
that can provide guarantees on the satisfaction of logical specifications by a
planner [61, 45]. Such tools could be used to provide fairness guarantees in the
form of a probability of satisfying a fairness constraint, bounds on the distance
840 between a predicted distribution and desired distribution of state features, etc.
However, how to efficiently satisfy non-Markovian fairness requirements with
such methods is still an open problem, especially since most of these methods

do not scale well with the size of the state space.

Sub-optimality and approximation guarantees are also important: users need
845 to be confident on the degree of error in fairness computations and the path
optimality. While some planning methods already provide sub-optimality guar-
antees [47], they also require the use of cumulative cost functions. An analysis
of error and sub-optimality bounds introduced by cumulative approximations of
the original cost functions will also be a necessary part of responsible algorithm
850 development.

5.2.5. Fairness: Including realistic fairness models within the planning objectives

Fairness with respect to locations cannot be fulfilled if there are missing
locations in a map, and similarly for fairness on protected characteristics. For
855 fair navigation planning systems to be reliable, data and models of the relevant
fairness features will be required. For example, population statistics over a
map, or accuracy of prediction algorithms and their biases should be available
to the planning algorithms. Some of these data already exist: census data
can be detailed in some countries, and many state-of-the-art machine learning
860 techniques are open-source. Some data do not currently exist: for example,
usage and population statistics of hospital corridors and rooms, computer vision
datasets with annotated personal characteristics, etc. The performance of fair
navigation planning algorithms will strongly depend on the quality of the data
and models used during the planning process, and these will be a crucial part
865 in the development of any such algorithm.

As we have seen, good resources of fairness models and formal specifications
are important to be able to operationalize fairness as well as anticipate issues
and design choices early in the development stage.

5.3. Research agenda

870 The previous discussion implicitly defines a research agenda for fair naviga-
tion planners in particular but also for conceptualizing fairness, and for respon-

sible innovation methods.

For better transparency of autonomous systems in general, and mobile robots in particular, we need methods for statistically-sound detection of inequality, as well as visualization of inequality and its trade-offs. For better human control, we need methods for combining user input with traditional motion planners. These should probably provide not only a fairness “knob” but also allow designers and users to anticipate issues, add more considerations and experiment with design options and simulated outcomes to increase an understanding of the problem. We need methods for privacy-preserving navigation planning [52], since guaranteeing fairness might require access to private information as we have discussed. For optimality and verifiability we need methods for either globally optimal solutions or (sub-) optimality guarantees on non-Markovian costs and constraints.

There are also social questions to ask about the fair design of mobile robots and autonomous systems. There is a need to investigate different ways to present robot deployment simulation results (such as the inequality plots we use in Figure 2), not merely in terms of static representations but in ways that animate the experiment, for example through walkthroughs [62], simulations [63] or quasi-naturalistic experiments [64, 65]. Indeed, the use case considered here suggests a novel programme of work involving a collaboration between social scientists and artificial intelligence researchers. For a number of years, social scientists have contributed to system development by undertaking distinct activities such as eliciting requirements and values from users, typically drawing from well-established methods such as semi-structured interviews and focus groups. With complex systems, such methods of elicitation may be problematic, as potential users and stakeholders have few resources to be able to assess putative technologies and their consequences. It may need a closer collaboration between social scientists and computer scientists to develop new methods for elicitation and engagement that allow participants to reason about and discuss the consequences of emerging technologies. This requires innovation in both computer science and social science methods.

This paper also raises fundamental philosophical questions about the ethical design of autonomous systems. What are the boundaries of what humans and machines should be responsible for in the design and execution of autonomous systems? When is affirmative action through robot behaviour acceptable? Does the “protected characteristic” framework make sense given the arbitrariness of the “characteristics” and the (sub-)categories that are chosen for the characteristics? For example, the census of the Office for National Statistics in England includes “white British”, “white Irish”, “white Gypsy”, “white other” as separate ethnicity categories, which raises questions of conceptual borders and the reasonableness of being fair to these specific categories (e.g. does demographic parity over these specific categories even make sense?). Finally, what does the field of meta-ethics have to say about algorithmic fairness and what can it contribute with operationally?

5.4. Methodological contribution

In this paper we have taken a different approach to considering ethical concerns with respect to fairness. Rather than set out a set of principles or raising a set of questions regarding equality or distribution, we have sought to examine fairness by unpacking the definitions and seeking to characterize fairness further. This has not been undertaken with the aim of seeking a sole, fixed, definition, or to stipulate what is “meant” by fairness, but rather to serve as a resource for development of systems. In the case of robot motion developing formal definitions and simulations reveals concrete examples of where certain decisions would be indirectly discriminatory. The examples we provide also suggest the design decisions that need to be made, for example with respect to the trade-offs with privacy or with efficiency. Hence, the development of a model could be seen as a useful method for anticipating the risk and consequences of an innovation, consistent with the principle of anticipatory governance in RI. However, rather than only reflecting on general principles of fairness or requesting stakeholders to anticipate the consequences of an innovation, the formal models and simulation-based investigation provide a more solid foundation on which to

initiate discussions that can anticipate the consequences of an innovation. So, in the case of robot navigation it is possible to provide examples of particular
935 decisions made in the design of an algorithm and their consequences which can be a resource in stakeholder workshops where potential users, developers, policy makers and members of the general public seek to anticipate the implications of a technological innovation. Modelling, formal specification and simulation can help provide a more systematic and informed foundation to such discussions,
940 prior to any development taking place.

Responsible innovation is an approach that is increasingly becoming embedded within the processes for the procurement of research. Although the approach addresses all kinds of innovation it has tended to focus on research that is at its earliest stages. However, recently, particularly through initiatives
945 by research funders seeking to mitigate against the risks of technological innovation, there have been measures to address a wider range of developments, particularly where it is not immediately apparent that there are ethical or societal concerns. The case considered in this paper reflects this progression: a seemingly trivial or mundane matter such as how a robot moves and navigates
950 through an environment can invoke ethical concerns, such as those relating to fairness. To meet the requirements of responsible innovation, developers need to engage with potential stakeholders. Typically, this is done through activities such as workshops where ethical and societal concerns are discussed. However, it is a challenge for stakeholders to anticipate the nature and consequences of
955 a technology that has not been designed, let alone implemented. We require novel methods from both technology innovators and social scientists to support this engagement. Novel ways of providing resources for anticipating the consequences of a new technology, such as those through formal specification, modelling and simulation may serve as methods for providing resources for such
960 collaborative and participative activities.

6. Conclusion

In this paper we explored the concept of fairness in the seemingly mundane, value-neutral, technical problem of robot navigation. We showed that there is a fairness dimension to robot navigation, using a walkthrough example of a rescue robot and brief comparisons to an autonomous vehicle and hospital robot use case. As we have discussed, the way robots move naturally changes the likelihood that certain people have of benefiting from access to the robot. This inequality, as well as structural inequalities (e.g. age-, race- and income-related spatial segregations in urban areas) can give rise to concerns of distributive justice. We discussed how mobile robots will have to face similar dilemmas that humans already face and how the notion of fairness will depend on the context. We then sought to build a resource for developers, of formal definitions, socio-technical challenges and design choices that have to be thought through when implementing fair navigation planners. We defined two kinds of fairness objects in the context of robot motion: state-visits and state-features. The first deals with being fair to locations in terms of the number of times they are visited. The second deals with visiting locations in a way that achieves a desired distribution of protected characteristics of the people along them (e.g. age, income). We then applied multiple theories of distributive justice to our navigation problem and obtained a collection of formal definitions. These definitions can be used to simulate robot deployment outcomes, ground discussions of fairness and design across multiple stakeholders, and anticipate issues. We showed that fairness-aware navigation planners will involve efficiency-fairness trade-offs, that their design should be an iterative process of understanding the fairness issues of the context at hand, and that current planning methods have downsides that need to be addressed (i.e. non-Markovian or lack of formal guarantees).

The methodology of the paper was to focus on a particular system to bring out these design choices and challenges through deployment simulations. We discussed the requirements for responsible design of such systems and the adequacy of the approach as a general method of Responsible Innovation that

anticipates ethical issues such as fairness and helps ground discussions and understanding of the problem by stakeholders.

This paper also sets the ground for a new research field of fair planning. Several challenges still lie ahead. Part of those are technical challenges of designing efficient, interpretable, formally verifiable and optimal methods to solve
995 non-Markovian fairness-aware planning problems. Another part is related to responsible innovation and value-sensitive design through appropriate analysis and visualization tools. Finally, the ethics of risk imposition [66], and the social and ethical implications of apparently innocent behaviour of autonomous
1000 systems are important topics to further explore. Such reflections are necessary for us to better align our robots with our values, and to better anticipate the impact autonomous systems have in society and our lives.

7. Appendix

7.1. The rescue-robot walkthrough example

Our analysis uses the openly-available data of the 2011 census of the Office
1005 for National Statistics in England, which maps over 400 variables among which population density and age distribution. We used the scripts from [67] to collect this data. In this paper we focus on 3 variables: age, ethnicity and gender.

It is unclear in the dataset which reasoning was used for selecting the categories.
1010 For example, the age categories are not uniform (some categories encompass only a couple of years, e.g. 16-17, while others are much larger, e.g. 75-84).

We locate the hypothetical base station of the robot in the city centre’s fire station (Rewley Road Fire Station). To simulate a thorough planner that starts
1015 exploring from the base station, we use a square region of 400 cells (approximately 1km²).

7.2. A fairness-aware human-in-the-loop planner

We consider a motion planning problem on a state-space \mathcal{S} and paths $\zeta = s_1, \dots, s_N$ where $s_i \in \mathcal{S}$, $\forall_{i=1, \dots, N}$. We are interested in optimizing both an

1020 efficiency-related function $f_{\text{efficiency}}(\zeta)$ (e.g. total population found along the path) and a fairness-related objective function $f_{\text{unfairness}}(\zeta)$.

We assume formal specification of fairness of the “affirmative action” type (Section 4.3), although the algorithm could equally be applied to any of the other specifications. We further assume a user or stakeholder provides a reference
1025 (desired) distribution Q for the characteristic of interest. In our experiments we set $Q = p_A$ (i.e. the city-wide distribution of age), which is equivalent to demographic parity. We compute the degree of (un)fairness as the Jensen-Shannon (JS) distance between the path-accumulated characteristic distribution p_A^ζ and the reference distribution Q :

$$f_{\text{unfairness}}(\zeta) := f_{\text{JS}}(\zeta) = \sqrt{D_{\text{JS}}(p_A^\zeta || Q)}, \quad (3)$$

1030 where

$$D_{\text{JS}}(P || Q) = \frac{1}{2} (D_{\text{KL}}(P || Q) + D_{\text{KL}}(Q || P)), \quad (4)$$

and where $D_{\text{KL}}(P || Q)$ is the Kullback-Leibler divergence between P and Q . The JS distance will be 0 if the class distribution along the path is exactly as desired, and 1 if the exact opposite.

To estimate trade-offs we use Pareto-front estimation in a multi-objective
1035 setting. An example Pareto-front is shown in Figure 3. It consists of a set of points that cannot be improved upon in one objective without making the other objective worse off. The use of Pareto-fronts allows to directly optimize fairness and efficiency trade-offs, while at the same time providing visualizations for problem intuition. Pareto-front estimation also provides a degree of human
1040 control over the algorithm, since each point of the Pareto-set will correspond to a different path, and a user is free to choose the one that leads to the most appropriate fairness-efficiency trade-off characteristics. Users can inclusively use the degree of trade-off together with other relevant expert knowledge to decide which is the best solution, or to revise the constraints on the planner (e.g. add
1045 an intermediate goal or change the location of the base station). It is in this sense that we call this a “human-in-the-loop” algorithm. We also propose to

visualize path-wise and desired distributions together with planned paths as a way to inform and potentially allow to discover fairness issues (see Figure 2).

Pseudo-code of our planner design is shown below in Algorithms 1 and 2. In simple words, we start by optimizing the waypoints of a robot trajectory

Algorithm 1 Fairness-aware human-in-the-loop planner

Input: state space \mathcal{S} , start state s_{start} , goal state s_{goal} , number of waypoints W , desired distribution Q , evolutionary optimization’s population size M
 // We will use Z to represent a vector of M waypoint-based paths
for Individual $i \leftarrow 1, \dots, M$ **do**
 $Z_i \leftarrow \text{UNIFORM_SAMPLES}(\mathcal{S}, W)$ ▷ Initialize each path as random waypoints
end for
 $F \leftarrow \text{EVALUATE}(Z, s_{\text{start}}, s_{\text{goal}}, Q)$
 $Z, F \leftarrow \text{PARETO_OPTIMIZE}(Z, F, \text{EVALUATE}())$
 // Display Pareto-front F to user
 // User chooses trade-off F_u (solution Z_u)
 $Z_{\text{full}} \leftarrow [s_{\text{start}}, Z_u, s_{\text{goal}}]$ ▷ Concatenate waypoints with start and goal
 $\zeta \leftarrow \text{INTERPOLATE}(Z_{\text{full}})$ ▷ fine interpolation between waypoints.
Output: ζ ▷ final path

1050

using a multi-objective evolutionary method for Pareto-front estimation [59]. The waypoints are connected to start and goal states, interpolated, and used to compute fairness and efficiency objectives. Then, the user visualizes the Pareto-front, possibly together with auxiliary information on each of the solutions along the front, such as the robot paths themselves, and the path-wise distributions.

1055

In deployment situations, a user can select one of the solutions along the Pareto-front, which is sent to the robot for execution.

To avoid conflating algorithm- with function-approximation issues, we used the same optimization method to optimize population density and the cumulative fairness cost (2) in Figure 5.

1060

Algorithm 2 EVALUATE

Input: set of paths $Z = Z_1, \dots, Z_M$, where $Z_i \in \mathcal{S}^W$, and W is the number of path waypoints, start state s_{start} , goal state s_{goal} , desired distribution Q

for Individual $i \leftarrow 1, \dots, M$ **do**

$Z_{\text{full}} \leftarrow [s_{\text{start}}, Z_i, s_{\text{goal}}]$ \triangleright Concatenate waypoints with start and goal

$\zeta \leftarrow \text{INTERPOLATE}(Z_{\text{full}})$ \triangleright fine interpolation between waypoints.

$E \leftarrow \text{EFFICIENCY_OBJECTIVE}(\zeta)$

$p_A^\zeta \leftarrow \text{COMPUTE_DISTRIBUTION}(\zeta)$

$J \leftarrow \text{JS_DISTANCE}(p_A^\zeta, Q)$

$F_i \leftarrow [E, J]$ \triangleright values of the objectives for Z_i

end for

Output: F \triangleright values of the objectives for all paths

Acknowledgements

We thank John Danaher for the comments and suggestions to an early version of this paper. This project was funded by the UK’s Engineering and Physical Sciences Research Council (EPSRC) through projects RoboTIPS (EP/S005099/1) and THuMP (EP/R033722/1).

References

- [1] A. Holzapfel, B. Sturm, M. Coeckelbergh, Ethical dimensions of music information retrieval technology, Transactions of the International Society for Music Information Retrieval (2018).
- [2] A. Golub, R. A. Marcantonio, T. W. Sanchez, Race, space, and struggles for mobility: Transportation impacts on african americans in oakland and the east bay, Urban Geography 34 (5) (2013) 699–728 (2013).
- [3] P. Lin, Why Ethics Matters for Autonomous Cars, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015, pp. 69–85 (2015). doi:10.1007/978-3-662-45854-9_4.

- [4] A. F. T. Winfield, M. Jirotko, Ethical governance is essential to building trust in robotics and artificial intelligence systems, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133) (2018) 20180085 (2018). doi:10.1098/rsta.2018.0085.
- 1080 [5] A. Winfield, An updated round up of ethical principles of robotics and ai (2019).
URL <https://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html>
- 1085 [6] R. Owen, P. Macnaghten, J. Stilgoe, Responsible research and innovation: From science in society to science for society, with society, *Science and Public Policy* 39 (6) (2012) 751–760 (12 2012). doi:10.1093/scipol/scs093.
- 1090 [7] R. Von Schomberg, A vision of responsible research and innovation, *Responsible innovation: Managing the responsible emergence of science and innovation in society* (2013) 51–74 (2013).
- 1095 [8] B. Grimpe, M. Hartswood, M. Jirotko, Towards a closer dialogue between policy and practice: Responsible design in hci, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14*, ACM, New York, NY, USA, 2014, pp. 2965–2974 (2014). doi:10.1145/2556288.2557364.
- [9] H. Webb, A. Koene, M. Patel, E. P. Vallejos, Multi-stakeholder dialogue for policy recommendations on algorithmic fairness, in: *Proceedings of the 9th International Conference on Social Media and Society, SMSociety '18*, ACM, New York, NY, USA, 2018, pp. 395–399 (2018). doi:10.1145/3217804.3217952.
- 1100 [10] H. Webb, M. Patel, M. Rovatsos, A. Davoust, S. Ceppi, A. Koene, L. Dowthwaite, V. Portillo, M. Jirotko, M. Cano, it would be pretty immoral to choose a random algorithm opening up algorithmic interpretability and transparency, in: *ETHICOMP 2018*, 2018 (2018).

- 1105 [11] W. Dieterich, C. Mendoza, T. Brennan, Compas risk scales: Demonstrating accuracy equity and predictive parity, Northpoint Inc (2016).
- [12] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias: theres software used across the country to predict future criminals. and its biased against blacks. propublica 2016 (2016).
- 1110 [13] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, *Big Data* 5 (2) (2017) 153–163 (2017). doi:10.1089/big.2016.0047.
- [14] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Trans-
1115 parency, Vol. 81 of Proceedings of Machine Learning Research, PMLR, New York, NY, USA, 2018, pp. 77–91 (23–24 Feb 2018).
- [15] B. Taati, S. Zhao, A. B. Ashraf, A. Asgarian, M. E. Browne, K. M. Prkachin, A. Mihailidis, T. Hadjistavropoulos, Algorithmic bias in clinical populationsevaluating and improving facial analysis technology in
1120 older adults with dementia, *IEEE Access* 7 (2019) 25527–25534 (2019). doi:10.1109/ACCESS.2019.2900022.
- [16] M. Brandao, Age and gender bias in pedestrian detection algorithms, in: Workshop on Fairness Accountability Transparency and Ethics in
1125 Computer Vision, CVPR, 2019 (Jun 2019).
URL <http://www.martimbrandao.com/papers/Brandao2019-fatecv.pdf>
- [17] G. Walker, Environmental justice: concepts, evidence and politics, Routledge, 2012 (2012).
- 1130 [18] T. Khaitan, Indirect discrimination, in: The Routledge Handbook of the Ethics of Discrimination, Routledge, 2017, pp. 30–41 (2017).

- [19] M. Hardt, E. Price, N. Srebro, et al., Equality of opportunity in supervised learning, in: *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323 (2016).
- 1135 [20] B. Eidelson, *Discrimination and disrespect*, Oxford University Press, 2015 (2015).
- [21] I. Hirose, *Egalitarianism*, Routledge, 2014 (2014).
- [22] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: *International Conference on Machine Learning*, 2013, pp. 325–333 (2013).
- 1140 [23] R. Zhang, M. Pavone, Control of robotic mobility-on-demand systems: a queueing-theoretical perspective, *The International Journal of Robotics Research* 35 (1-3) (2016) 186–203 (2016).
- [24] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ACM, 2012, pp. 214–226 (2012).
- 1145 [25] R. Dworkin, What is equality? part 2: Equality of resources, *Philosophy & public affairs* (1981) 283–345 (1981).
- [26] J. Rawls, *A Theory of Justice*, Harvard University Press, 1971 (1971).
- 1150 [27] H. Heidari, M. Loi, K. P. Gummadi, A. Krause, A moral framework for understanding of fair ml through economic models of equality of opportunity, *arXiv preprint arXiv:1809.03400* (2018).
- [28] J. P. Bigham, Bias and fairness in ml systems for people with disabilities (2019).
- 1155 URL <http://jeffreybigham.com/blog/2019/bias-and-fairness-in-ml-systems-for-people-with-disabilities.html>
- [29] J. Skorupski, *Ethical Explorations*, Oxford University Press, 1999 (1999).

- [30] D. Parfit, Equality or priority? in the ideal of equality, ed. matthew clayton
1160 and andrew williams (2000).
- [31] O. F. Norheim, A note on brock: prioritarianism, egalitarianism and the
distribution of life years, *Journal of Medical Ethics* 35 (9) (2009) 565–569
(2009).
- [32] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair
1165 determination of risk scores, arXiv preprint arXiv:1609.05807 (2016).
- [33] K. J. Arrow, A difficulty in the concept of social welfare, *Journal of political
economy* 58 (4) (1950) 328–346 (1950).
- [34] A. Narayanan, Translation tutorial: 21 fairness definitions and their poli-
tics, in: *Proc. Conf. Fairness Accountability Transp.*, New York, USA, 2018
1170 (2018).
- [35] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic de-
cision making and the cost of fairness, in: *Proceedings of the 23rd ACM
SIGKDD International Conference on Knowledge Discovery and Data Min-
ing*, ACM, 2017, pp. 797–806 (2017).
- 1175 [36] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, H. Wallach, A re-
ductions approach to fair classification, arXiv preprint arXiv:1803.02453
(2018).
- [37] D. P. OMathúna, B. Gordijn, M. Clarke, Disaster bioethics: Normative
issues when nothing is normal, Vol. 2, Springer Science & Business Media,
1180 2013 (2013).
- [38] O. Merin, N. Ash, G. Levy, M. J. Schwaber, Y. Kreiss, The israeli field
hospital in haitiethical dilemmas in early disaster response, *New England
Journal of Medicine* 362 (11) (2010) e38 (2010).
- [39] G. Bognar, I. Hirose, *The Ethics of Health Care Rationing: An Introduc-
tion*, Routledge, 2014 (2014).
1185

- [40] R. Avraham, Discrimination and insurance, The Routledge Handbook To Discrimination Lippert-Rasmussen Ed (2017).
- [41] G. E. Fainekos, H. Kress-Gazit, G. J. Pappas, Hybrid controllers for path planning: A temporal logic approach, in: Proceedings of the 44th IEEE Conference on Decision and Control, IEEE, 2005, pp. 4885–4890 (2005).
1190
- [42] M. Cirillo, F. Pecora, H. Andreasson, T. Uras, S. Koenig, Integrated motion planning and coordination for industrial vehicles, in: Twenty-Fourth International Conference on Automated Planning and Scheduling, 2014 (2014).
- [43] E. Plaku, Path planning with probabilistic roadmaps and co-safe linear temporal logic, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2012, pp. 2269–2275 (2012).
1195
- [44] G. De Giacomo, F. Patrizi, S. Sardina, Generalized planning with loops under strong fairness constraints, in: Twelfth International Conference on the Principles of Knowledge Representation and Reasoning, 2010 (2010).
1200
- [45] B. Lacerda, D. Parker, N. Hawes, Multi-objective policy generation for mobile robots under probabilistic time-bounded guarantees, in: Proc. of the 27th Int. Conf on Automated Planning and Scheduling (ICAPS), Pittsburgh, PA, USA, 2017 (2017).
- [46] T. Khaitan, S. Steel, Wrongs, group disadvantage, and the legitimacy of indirect discrimination law, Bloomsbury Professional, 2017 (2017).
1205
- [47] M. Likhachev, G. J. Gordon, S. Thrun, Ara*: Anytime a* with provable bounds on sub-optimality, in: Advances in Neural Information Processing Systems, 2003, pp. 767–774 (2003).
- [48] S. Karaman, E. Frazzoli, Incremental sampling-based algorithms for optimal motion planning, Robotics Science and Systems VI 104 (2) (2010).
1210

- [49] B. Friedman, P. H. Kahn, A. Borning, A. Huldtgren, Value sensitive design and information systems, in: Early engagement and new technologies: Opening up the laboratory, Springer, 2013, pp. 55–95 (2013).
- 1215 [50] J. C. G. Higuera, A. Xu, F. Shkurti, G. Dudek, Socially-driven collective path planning for robot missions, in: 2012 Ninth Conference on Computer and Robot Vision, IEEE, 2012, pp. 417–424 (2012).
- [51] L. Boloni, S. A. Khan, S. Arif, Robots in crowds being useful while staying out of trouble, in: Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013 (2013).
- 1220 [52] A. Prorok, V. Kumar, Privacy-preserving vehicle assignment for mobility-on-demand systems, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 1869–1876 (Sept 2017). doi:10.1109/IROS.2017.8206003.
- 1225 [53] S. Feng, E. Whitman, X. Xinjilefu, C. G. Atkeson, Optimization-based full body control for the darpa robotics challenge, Journal of Field Robotics 32 (2) (2015) 293–312 (2015).
- [54] S. Caron, A. Kheddar, Multi-contact walking pattern generation based on model preview control of 3d com accelerations, in: Proceedings of the 2016 IEEE-RAS International Conference on Humanoid Robots, IEEE, 2016 (Nov. 2016). doi:10.1109/HUMANOIDS.2016.7803329.
- 1230 URL <https://hal.archives-ouvertes.fr/hal-01349880>
- [55] A. Lavin, A pareto optimal d* search algorithm for multiobjective path planning, arXiv preprint arXiv:1511.00787 (2015).
- 1235 [56] S. Choudhury, C. M. Dellin, S. S. Srinivasa, Pareto-optimal search over configuration space beliefs for anytime motion planning, in: Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on, IEEE, 2016, pp. 3742–3749 (2016).

- [57] A. Critch, Toward negotiable reinforcement learning: shifting priorities in pareto optimal sequential decision-making, arXiv preprint arXiv:1701.01302 (2017).
- [58] A.-I. Mouaddib, Vector-value markov decision process for multi-objective stochastic path planning, *International Journal of Hybrid Intelligent Systems* 9 (1) (2012) 45–60 (2012).
- [59] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, *IEEE Transactions on Evolutionary Computation* 6 (2) (2002) 182–197 (Apr 2002). doi:10.1109/4235.996017.
- [60] E. Zitzler, M. Laumanns, L. Thiele, Spea2: Improving the strength pareto evolutionary algorithm, TIK-report 103 (2001).
- [61] N. Hawes, C. Burbridge, F. Jovan, L. Kunze, B. Lacerda, L. Mudrova, J. Young, J. Wyatt, D. Hebesberger, T. Kortner, et al., The strands project: Long-term autonomy in everyday environments, *IEEE Robotics & Automation Magazine* 24 (3) (2017) 146–156 (2017).
- [62] P. G. Polson, C. Lewis, J. Rieman, C. Wharton, Cognitive walkthroughs: a method for theory-based evaluation of user interfaces, *International Journal of man-machine studies* 36 (5) (1992) 741–773 (1992).
- [63] J. F. Kelley, An iterative design methodology for user-friendly natural language office information applications, *ACM Transactions on Information Systems (TOIS)* 2 (1) (1984) 26–41 (1984).
- [64] L. A. Suchman, *Plans and situated actions: The problem of human-machine communication*, Cambridge university press, 1987 (1987).
- [65] C. Heath, P. Luff, The naturalistic experiment: Video and organizational interaction, *Organizational Research Methods* 21 (2) (2018) 466–488 (2018). doi:10.1177/1094428117747688.

- 1265 [66] J. Oberdiek, Imposing risk: a normative framework, Oxford University Press, 2017 (2017).
- [67] A. Singleton, 2011 census open atlas. university of liverpool: Consumer data research centre (cdrc) (2015).
URL <https://data.cdrc.ac.uk/product/cdrc-2011-census-open-atlas>
- 1270