The Broomean Fairness Measure for Queueing

Martin Nørgaard Petersen
Erasmus Institute for Philosophy of Economics (EIPE)
Erasmus University Rotterdam

Supervised by:
Dr. Conrad Heilmann

Advised by:
Dr. Stefan Wintein

June 2024

Abstract

I present a new measure for assessing the fairness of queues: The Broomean Fairness Measure for Queueing (BFMQ). The measure is based upon the adaptation of Broomean fairness from Wintein and Heilmann (2024a; 2024b), according to which fairness has both an absolute and a comparative dimension. I provide a conceptual reconstruction of fairness for queueing, arguing that certain queueing problems can be considered problems of the fair division of a limited resource. Accordingly, I introduce and suggest possible implementations of fairness 'claims' with associated 'amounts' and 'strengths', for queueing. Finally, I propose a method for measuring the fairness of different queueing scenarios which generalises and improves on the Resource Allocation Queueing Fairness Measure of Raz, Levy and Avi-Itzhak (2004).

This thesis is presented towards to the degree as Research Master in Philosophy and Economics at the Erasmus University Rotterdam.

Contents

Word count: 19.850 (excluding Contents, References and Appendices)

1 Introduction

Queues are everywhere in daily life: when waiting at the cashier in the supermarket, when waiting for treatment at the hospital, when asking your computer to run a large number of jobs, or when trying to buy tickets to your next Taylor Swift concert. Whenever the demand for service exceeds the available server capacity, customers can either be refused service or — more naturally — be asked to wait in line. And suddenly, a queue has formed.

It is common to let people be served in the order they arrive. This *queueing discipline* is known as First-Come-First-Serve. However, other disciplines exist. For instance, many airlines board families with children before other passengers, or offer passengers to pay to board before others. And certain ticket vendors for popular concerts serve people at random. While queues in everyday life are often characterised by service being provided to one customer at a time, there is a wide body of queues, especially in computer science, where more or all customers are served simultaneously, albeit at a slower rate.

Most often, work on queues deals with matters regarding effectiveness: how can the total waiting time of the people queueing be reduced? The concern for effectiveness is often founded in economic interests: if overall waiting time in society is reduced, less resources are wasted. But a one-sided focus on effectiveness may have adverse effects for other concerns, such as fairness. When regarding waiting time only, one may find that it is optimal to let some people wait for longer than what their claim to service otherwise dictates. In other words, we may choose to systematically assign — on grounds of effectiveness — a larger part of our resources to some particular individuals than what they have a claim to. This is unfair.

It doesn't follow that because a queueing scenario is not fair, we necessarily ought to do something about it. It might be that effectiveness overrules all other concerns, or that fairness is simply not relevant in the particular setting. But there are scenarios were fairness *does* matter. And in these cases it is necessary to assess and compare the fairness of different queueing disciplines, if we want to make improvements.

Therefore, I propose a measure for determining the fairness of queueing disciplines. Previous attempts have been made at developing measures of the fairness of queueing disciplines. These measures, however, often lack rigour with regards to the use of philosophically loaded terms, and in particular with regards to their definition of fairness. In contrast, the measure I propose differentiates itself from other measures in the literature by an uncompromising foundation in a philosophical account of fairness.

I take my outset in The Resource Allocation Queueing Fairness Measure (RAQFM) of Raz, Levy and Avi-Itzhak (2004). I do so for two reasons. First, it possesses a well-developed technical apparatus. And second, it models queueing as the allocation of a scarce resource. Regarding queueing as a matter of dividing a scarce resource makes it possible to actuate the well-developed philosophical literature on far division problems. By employing this literature, my measure improves on measures such as the RAQFM which, despite interpreting queueing as a fair division problem, does not engage with the relevant literature on the topic.

1.1 Roadmap

In the following section, I introduce fairness measures for queueing and present the RAQFM. In Section 3, I discuss queueing effectiveness and show how it sometimes conflicts with fairness. I explain why and when fairness is relevant to consider when working with queues. Section 4 provides a detailed and general definition of fairness. On basis hereof, in Section 5, I conceptualise fairness in queueing and argue that the problem of queueing can be considered a problem of fair division. I also work out the details for representing this fair division problem as a *fairness structure*. Taken together, these parts provide *the foundation* for the Broomean Fairness Measure for Queueing (BFMQ).

Section 6 provides examples of the application of the Absolute Priority Rule of Wintein and Heilmann (2024a). Given the conceptual foundation, Section 7 explains how the fairness of queues can be measured which ultimately culminates in the introduction of the BFMQ. Section 8 concludes.

2 Fairness Measures for Queueing

In this section, I introduce and discuss the Resource Allocation Queueing Fairness Measure (RAQFM) of Raz, Levy and Avi-Itzhak (2004). This serves two purposes. First, it introduces the notion of a fairness measure for queueing and thus provides an idea of the end goal of this project. Second, my Broomean Fairness Measure for Queueing (BFMQ) has a number of elements in common with the RAQFM, while at the same time improving on the RAQFM by relying on a rigorous and precise conceptualisation of fairness for queueing. Introducing the RAQFM therefore provides a good starting point for the further development.

In effect, this section attaches the BFMQ to the existing literature and lays out the aspects on which it seeks to improve current measures. But before examining the RAQFM in detail, let me give a few introductory remarks on queueing.

2.1 A primer to queueing

Queues can be observed across multiple domains where they are commonly used to schedule and prioritise between different tasks. Queues typically arise due to constraints on the server's capacity: whenever there are physical or economical constraints that imply that clients cannot be served instantly, a natural solution is to use queues (as opposed to simply refusing service to clients). Most people experience queues every day, for instance when waiting in line at the supermarket. But queueing is also widely utilised on e.g., computer networks for controlling the flow of communication.

When formalised, a queueing system is commonly described by three characteristics: the input process, the service mechanism and the queue discipline (Cooper 1981, 73f.). First, the *input process* describes the sequence of requests for service and can be described in terms of the time between consecutive clients' arrival instants. Second, the *service mechanism* describes how clients are served. This includes the *number of servers* and the length of time that the client occupies the server, also known as the *service requirement* (Avi-Itzhak, Levy and Raz 2008, p. 498). When modelling queues, the input process and the service mechanism are often assumed to be stochastic processes, meaning

that one cannot predict individual arrivals, or how long it takes to serve a given client, except in probabilistic terms.

Knowing the input process and the service mechanism, one can determine the likelihood that *blocking* occurs which is the option that a client arrives and there are no servers idle. The third characteristic determining queueing systems is thus how such blocked clients are treated. The *queue discipline* thus determines the disposition of clients that are not served immediately.

Queueing systems can vary along any of these dimensions, yielding a large variety of queues. As a result, one should not conflate queueing with 'waiting in line'. While the First-Come-First-Serve (FCFS) discipline, where clients are served in order of their arrival, is commonly observed in physical queues, other disciplines that allow for serving more or all customers at the same time at a slower rate are prevalent in other applications. As I explain later, the subset of queueing systems that allow for simultaneous service to all clients plays a central role for this project.

2.2 The Resource Allocation Queueing Fairness Measure

The Resource Allocation Queueing Fairness Measure (RAQFM) of Raz, Levy and Avi-Itzhak (2004) (and Avi-Itzhak, Levy and Raz (2007)) is a measure for quantifying the fairness of the realised disposition of clients in a queue. The RAQFM applies to queues generally, both where service is provided in a one-by-one manner, and where service is provided to all customers simultaneously. Notably, the RAQFM relies on the assumption that queueing can be interpreted as the allocation of a scarce resource, where the resource to be allocated is the capacity of a server. There are alternative ways of interpreting queueing for the purposes of creating fairness measures, and in Appendix A.1, I provide a brief overview.

The RAQFM considers the set $N(t)$ of clients for every epoch t, where the number of customers in the set is given by $n(t)$. At every epoch, there is also a 'warranted rate' of service $r_i(t)$ which is the rate of service that customer i deserves. Avi-Itzhak, Levy and Raz (2007, p. 66) argue that '[t]he fundamental principle underlying RAQFM is the belief that at every epoch t, all customers present in the system deserve an equal share of the system resources'. As such, the authors believe that there are no reasons for assigning a larger share to some customers rather than others, and hence they suggest that the warranted rate for all customers is given by $r_i(t) = r(t) = \frac{1}{n(t)}$.

The information relevant to the RAQFM, for a specific epoch t, can be summarised by the following tuple (this notation is new to the RAQFM, but shall prove convenient later, cf. Section 5.1):

$$\mathcal{Q}_t = \langle C, N, r \rangle_t = \left\langle 1, N, \frac{1}{n} \right\rangle_t , \tag{2.1}$$

where I've normalised server capacity $C = 1$ on the right hand side, for consistency with Raz, Levy and Avi-Itzhak (2004).

In a specific queueing scenario, the service actually provided to customer i at epoch t is denoted $s_i(t)$. Accordingly, Avi-Itzhak, Levy and Raz (2007) define 'individual discrimination' as:

$$d_i^{\text{RAQFM}}(t) \equiv s_i(t) - R(t) = s_i(t) - \frac{C}{n(t)}, \tag{2.2}$$

where the warranted amount is given as the product of the warranted rate and total server capacity[1], i.e., $R(t) = C \cdot r(t)$.

Individual discrimination is thus the difference between what the customer deserves and the resources that she actually gets allocated at every epoch. Avi-Itzhak, Levy and Raz (2007) are using desert and fairness synonymously. The assumption that all customers deserve an equal share, thus implies that a queueing discipline is considered fair if all customers are assigned equal shares of the resource. So according to Avi-Itzhak, Levy and Raz (2007), fairness demands that the resource actually allotted to an individual in every epoch is $s_i(t) = \frac{C}{n(t)}$, in which case individual discrimination is $d_i^{\text{RAQFM}}(t) = 0$.

Queueing disciplines describe the *intended* disposition of clients in a queue. In practice, however, the service actually provided can differ from the intended service. Hence, individual discrimination can arise from one of two sources: either because the realised allocation of resources deviates from the one recommended by the queueing discipline (e.g., due to feasibility constraints), or because the queueing discipline itself is unfair.

Disregarding feasibility constraints for a moment, the queueing discipline that implies $s_i(t) = \frac{C}{n(t)}$ is called *Processor Sharing*. Processor Sharing assumes that all customers can be served at the same time and recommends that every customer is assigned an equal share of the resource. Therefore, the authors conclude that Processor Sharing is *the* fair queueing discipline (Avi-Itzhak, Levy and Raz 2007, Theorem 3.3). Let me illustrate what this implies, by means of example.

Example 2.1. 'Four Mathematicians' Suppose there are four mathematicians in a queue who need a computer to compute some difficult math problem for them. The first mathematician, Almaviva, requires 600 units of computing power to compute his problem. The second mathematician, Barbarina, requires 450 units, Constanze requires 300 units, and Despina requires 150 units. The computer has a total capacity of 1000 units per second. The computer is not restricted to serving one mathematician at a time, but can serve all four mathematicians at the same time. This implies that the queueing problem (for the first second, $t = 1$) can be represented by the tuple:

$$\mathcal{Q}_1^{\text{FM}} = \langle 1000, \{A, B, C, D\}, (1/4, 1/4, 1/4, 1/4) \rangle_1$$

Assume that Processor Sharing is implemented as the queueing discipline. Processor Sharing recommends that every mathematician is allocated an equal amount of the computers capacity, i.e., $s_i(t) = 1000 \cdot \frac{1}{4} = 250$ units per second each. If there are no feasibility constraints, then $\forall i : d_i^{\text{RAQFM}}(1) = s_i(1) - R(1) = 250 - 250 = 0$

Individual discrimination denotes discrimination at a given epoch t. The *total* discrimination is computed by aggregating individual discrimination for the total time the customer is in the queue. If b_i denotes the arrival time and

[1]The reader may confirm that for $C = 1$, $R(t) = r(t)$.

d_i denotes the exit time for customer i, then Raz, Levy and Avi-Itzhak (2004) define the overall discrimination as:

$$D_i^{\text{RAQFM}} \equiv \sum_{i|t\in(b_i,d_i)} d_i^{\text{RAQFM}}(t) \tag{2.3}$$

An important property of total discrimination is that, whenever the server is allotting all of its resource, every positive discrimination is balanced by a negative discrimination. Allotting the resource is a zero-sum game, so to say: giving more to one customer implies taking it from another. As a result, the fairness of a queueing discipline can not be measured by average discrimination, as the average discrimination will be $\mathrm{E}(D_i^{\text{RAQFM}}) = 0$, whenever the server's capacity is exhausted. Instead, it is measured by the *dispersion* of the overall discrimination. Raz, Levy and Avi-Itzhak (2004) assume that *variance*, is the best operator for this purpose.

The fairness of a specific queueing scenario S, according to the Resource Allocation Queueing Fairness Measure, is thus given as:

$$\psi^S = \frac{1}{n}\sum_{i=1}^{n}\left(D_i^{\text{RAQFM}}\right)^2$$

A queueing *scenario* refers to the specific disposition of clients, i.e., an actual allotment of resources to clients. For queueing systems that enter steady state (statistical equilibrium) it is possible to assess the fairness of a particular queueing *system* and thus make comparisons of the fairness of different queueing disciplines. Given a fairness measure for queueing scenarios, providing an associated measure for queueing systems is only a matter of technicalities. I therefore neglect it in this thesis. I make brief mention of it in Appendix A.2 and otherwise refer to Avi-Itzhak, Levy and Raz (2007) and Raz, Levy and Avi-Itzhak (2004).

2.2.1 Raz, Levy and Avi-Itzhak's 'Principles of fairness'

To assess the validity of the RAQFM, Avi-Itzhak, Levy and Raz (2007) introduce two *principles of fairness* and examine the behaviour of the RAQFM — i.e., comparative values of fairness — in particular cases. According to Avi-Itzhak, Levy and Raz (2008), both principles are based on 'basic intuitions' and are defined as:

Seniority Preference Principle *(or temporal fairness)* It is fair that jobs are served according to arrival times.

Service Requirement Principle *(or proportional fairness)* It is fair for large jobs to have large response times in comparison to short jobs, or response times should be proportional to job size[2].

In conclusion, Avi-Itzhak, Levy and Raz (2007) argue that the RAQFM reacts in intuitive and proper ways as determined by these principles.

[2] Avi-Itzhak, Levy and Raz (2008, p. 498) express this as the principle that '*the one who demands the least of the server's time should be served first*'.

2.3 Shortcomings of the RAQFM

The RAQFM comes short on four points. First, Raz, Levy and Avi-Itzhak (2004) are assuming that all individuals deserve the same amount of the resource. I argue that this assumption is unfounded. Consider again the example of the four mathematicians. I can easily imagine a situation where one mathematician deserves more than the others. Maybe solving the problem of one of the mathematicians has very large consequences — it might solve the problems of climate change. Should she not deserve to use more of the computer's resources than the others? I believe so. Or, maybe solving the problem of another mathematician is very urgent — it might save someone's life, if solved. Should she not deserve more of the resource? I definitely think so. It seems to me that it is often the case that there are differences in individuals' claims to the resource. And regardless, a measure of fairness that does not rely on the assumption of equal claims to the resource, is more versatile and thus more attractive.

Second, the RAQFM does not factor in the individual requirements when determining the fairness of a queueing discipline. As a result, the queueing problem considered by the RAQFM (Eq. 2.1) does not entail the individual requirements. As I argue in Section 4.3.2, the mere fact that some individuals require more service than others may imply that their claim to the resource exceeds that of others. It also implies that some individuals may be assigned more of the resource than they require. In the example, one of the mathematicians requires 150 units to compute her problem but receives 250 units. It seems unfair that one individual receives more than she actually requires, while others are receiving less than they require, at the same time. The wasted capacity of the server could be allocated to one of the other mathematicians, without harming anyone. The definition of fairness, that I introduce in Section 4, takes into account this 'absolute' dimension of fairness.

Third — and somewhat more tacit — Raz, Levy and Avi-Itzhak (2004) suppose that the most appropriate measure of dispersion is variance. Variance assigns large values of discrimination a comparatively higher weight than small values. That is, the choice of dispersion operator caries some implicit relative valuation of different magnitudes of discrimination. The authors fail to recognise the philosophical commitments that follow with this choice of operator. In section 7.2, I lay out the philosophical commitments that follow different choices of dispersion operators and argue that *mean absolute deviation* is a better measure of dispersion than variance, for the purposes of a fairness measure.

Fourth, the principles of fairness of Raz, Levy and Avi-Itzhak (2004) are dubious. Notably, both principles of fairness are shared by many other scholars in the literature (see e.g., Wierman (2011) or Maccio, Hogg and Down (2018, p. 133)), so this point of criticism stretches further than the RAQFM. Although both seniority and service requirement are morally relevant to fairness, I argue in Section 5 that they are so in very different ways than how Raz, Levy and Avi-Itzhak (2004) suggest. One the one hand, I agree with the authors' basic intuition that higher seniority should contribute to priority for service. However, I suggest conceptualising it in a different way. On the other hand, I disagree with the principle that shorter jobs, should have shorter response time. To the contrary, it follows from my conceptualisation of fairness for queueing that individuals with jobs that require more of the resource, also have a larger claim to that resource.

2.4 A new measure with an improved philosophical foundation

All of the above issues can be rectified by introducing precise definitions of the philosophical terms used and by uncovering implicit philosophical commitments. Raz, Levy and Avi-Itzhak (2004) assume that queueing can be modelled as a resource allocation problem, while at the same time do not take the advantage of the literature of the fair allocation of scarce resources. Instead,Avi-Itzhak, Levy and Raz (2008, p. 503) suggest that fairness is an intuitive concept that is understood inherently by most people. For this reason, the measure of fairness they propose is 'not necessarily directly based on some abstract general formal conception offered by "deep thinkers."'. Although I agree that most people may have an intuitive sense for what is fair and what is not, I do not think that the same people could provide a principle that is sufficiently generalisable to be implemented in a measure of fairness. Without such principle it is difficult to make consistent assessments of which queueing disciplines are fair and which are not. So contrary to Avi-Itzhak, Levy and Raz (2008), I do believe there is room for some 'deep thinking'.

Raz, Levy and Avi-Itzhak (2004) — like other contributors to the literature — are using loaded terms with little regard for their definitions and interdependencies. For instance, Raz, Levy and Avi-Itzhak (2004) are implicitly assuming that it is *fair* if people get what they *deserve*. This is necessary for them to conclude that an allocation is fair whenever it allocates an equal amount of the resource to every individual, so long as everybody is equally deserving. But desert is not all that matters to fairness. In fact, I'm not at all certain about how fairness and desert are related. I briefly return to the point in Section 4.3.2 but will generally avoid using the term 'desert' for the remainder of the thesis and solely focus on 'fairness'.

Avi-Itzhak, Levy and Raz (2008, p. 498) acknowledge the difficulties of determining a proper definition of fairness. They admit that interpreting the literature on fairness is 'beyond [their] ability', which amongst other things comes to show from their conflation of 'fairness' and 'social justice' (see Avi-Itzhak, Levy and Raz 2008, p. 497). But aside from the weak conceptual foundation of their measure, it provides a good technical foundation.

The main aim and contribution of this paper is thus to reconstruct how fairness is conceptualised in queueing theory and to provide a refined definition of fairness in queueing. Importantly, I rely on a general account of fairness that has been developed separately from concerns of queueing. This serves to avoid the risk of 'contaminating' any axiomatic principles of fairness with the conclusions one seeks to obtain. This risk is left ubiquitous when the axiomatic principles are defined and applicable only in relation to the specific domain targeted. I believe the 'principles of fairness' cited above exemplify this worry (cf. Section 2.2.1).

Moreover, relying on a carefully defined account of fairness provides an added clarity as to what determines the reasons, with regards to fairness, for serving some customers before others, and serving some customers more than others. Specifically, I draw on the Broomean account of fairness and introduce the notion of claims. I explain how one can make sense of their respective amounts and strengths in queuing applications. As far as I'm aware, the introduction of claims, and especially the separation into claim amount and claim strength,

is new to the field of fair queueing and provides a conceptual advancement of fairness in queueing.

2.4.1 The Broomean Fairness Measure for Queueing

My reconstruction of fairness for queueing forms the basis for a new measure: The Broomean Fairness Measure for Queueing (BFMQ). I argue that the individual's requirement for service is relevant information to the representation of queueing problems for the purposes of determining the fair allocation of service. In turn, I make the case that certain queueing problems can be represented as 'Broomean problems'. I thereby actuate the literature on fair division of scarce resources and contend, in line with Wintein and Heilmann (2024a), that the ideal, fair allocation is determined by the Absolute Priority Rule $P^{\dagger}(\mathcal{B}(t))_i$. Thus, individual discrimination is measured as the difference between the resources actually allotted and the allocation recommended by this rule:

$$d_i(t) = s_i(t) - P^{\dagger}(\mathcal{B}(t))_i. \tag{2.4}$$

I argue that the mean absolute deviation is a more appropriate measure of the dispersion of total discrimination than variance. Accordingly, the fairness of a queueing scenario, as judged by the BMFQ, is given as:

$$\phi^S = \frac{1}{n}\sum_{i=1}^{n} |D_i|, \tag{2.5}$$

where total discrimination is defined analogous to Eq. 2.3.

In many regards, the BMFQ generalises and improves on the RAQFM. But by one aspect, the BMFQ is reductive. The conceptual work I do here only considers queues where customers can be served all at once. As a result the BMFQ only applies to this particular type of queues, whereas the RAQFM applies to a wider set of queueing systems. The reason is that the literature on the division of indivisible goods is still full of unresolved caveats, which I do not settle in this thesis. It is however my conviction that given the work done here, if an operationalisation of the fair division of indivisible goods is provided, the extension of the BMFQ to a wider set of queueing systems should follow with only few necessary amendments.

3 Effectiveness and Fairness in Queueing

By far the greatest part of work on queues in operations research is concerned with optimising the effectiveness of queues. Concerns over the fairness of queueing disciplines has had little traction in the literature. In this section, I discuss the relation between effectiveness and fairness of queues, and further motivate why and when fairness is a relevant concern.

3.1 Effectiveness versus fairness

The effectiveness of a queue expresses how effective, or quickly, a queueing system can serve a set of customers. Given constraints on resources, one queueing system is considered more effective than another if the average waiting time

is lower in one system than in the other[3]. Thus, optimising the waiting time of a queueing system means determining the queueing discipline that reduces the expected average waiting time of the system, given some constraint on the server's capacity.

It is possible to derive the general result that in order to reduce average response time (waiting time plus service time) as much as possible, one should serve the job with the shortest remaining service time first. This holds for preemptive disciplines, i.e. where the service of a client can be stopped and re-uptaken later, in single server systems (Maccio, Hogg and Down 2018, p. 140; Wierman 2011, p. 39). Among non-preemptive disciplines, it is similarly most effective to serve the client with the shortest service requirement first. This is also known as the 'Shortest-Job-First'-discipline. To see this, consider the following example.

Example 3.1. 'Mr. Short and Mr. Long' Imagine that two customers arrive at the cashier in a supermarket simultaneously. Mr. Long has two loaded carts whereas Mr. Short has only one item. Then it is most effective that Mr. Short is served before Mr. Long rather than vice versa. Suppose that it takes 2 minutes to process all the goods in the cart, but only 15 seconds to process the one item. Serving Mr. Long first would mean that the total response time for both customers would be 4 minutes and 15 seconds (2×2 minutes $+ 15$ seconds). Serving Mr. Short before Mr. Long would be 2 minutes and 30 seconds ($2 \times$ 15 seconds $+ 2$ minutes).

Implementing disciplines such as Shortest-Job-First, in order to optimise the effectiveness of a queueing system, implies that customers with small service requirements are treated first at the expense of customers with large service requirements. This is at odds with fairness, I argue.

On my definition, fairness requires that claims are satisfied in proportion to their strength[4]. Giving priority to short service requirements is therefore only fair if customers with small service requirements would have larger claims to service compared to customers with large service requirements. I find the existence of such inverse dependency, between service requirement and claims to service, hardly convincing. In fact, I find it so hard to believe that I will argue for the opposite dependency, i.e., proportionality between claim amount and the extent of service requirement, in Section 5.

So, effectiveness does not imply fairness. Notably, this conflicts with the *service requirement principle* of fairness by Avi-Itzhak and Levy (2004), cited in Section 2.2.1. Their principle suggests that it is *fair* that customers with large service requirements wait longer[5]. Avi-Itzhak, Levy and Raz (2008) argue in favour of the principle by referring to peoples' 'personal experience' from

[3]Actually, the effectiveness of a queueing system can be measured by a number of different measures. It may be measured by (1) the *time waited*, (2) the *number of clients waiting* or (3) the *number of servers idle* (Shortle et al. 2017, p. 3). For these purposes, it does not matter exactly how effectiveness is measured.

[4]I define fairness more elaborately in Section 4.2.

[5]Wierman (2011, p. 41) goes as far as suggesting that this principle is an instantiation of the Aristotelian notion of fairness: equals should be treated equally, and unequals unequally, in proportion to relevant similarities and differences (Aristotle (2009) as paraphrased in Wintein and Heilmann (2024b, p. 22)). He therefore denotes the principle 'proportional fairness' (not to be confused with my proportionality requirement of Section (4.4)). I disagree with Wierman's application of the Aristotelian notion of fairness.

shopping in supermarkets. They consider the above example of Mr. Short and Mr. Long and conclude that it is *fair* for Mr. Long to let Mr. Short pass, even if Mr. Short arrives shortly after Mr. Long. I do not agree.

I argue that Avi-Itzhak, Levy and Raz's reasoning is not rooted in fairness but in other concerns, such as effectiveness or other-regarding preferences. Mr. Long may let Mr. Short pass because he assesses that the costs to fairness by letting Mr. Long pass him are outweighed by the costs of waiting imposed on Mr. Short. These costs may matter to Mr. Long because he has preferences for the well-being of Mr. Short, i.e., other-regarding preferences. But Mr. Long does not let Mr. Short pass *because of* fairness. So although it might be completely *appropriate* for Mr. Long to let Mr. Short pass, it is not necessarily *fair*. As a principle of fairness, I therefore refute the 'service requirement principle' of Avi-Itzhak, Levy and Raz (2008).

By subscribing to the service requirement principle, Avi-Itzhak and Levy (2004) are implying that effectiveness is inherent to fairness. It is not, however. Effectiveness and fairness are separate notions, and therefore must be treated separately. Here, I treat fairness by itself and leave concerns of effectiveness to a side[6]. Decoupling fairness from effectiveness ultimately means that a queueing discipline can be assessed to be completely fair but very ineffective at the same time. Hence, we should not disregard the effectiveness of queueing altogether, even though that for these purposes I will do so.

3.2 Effectiveness and efficiency

I am using the term *effectiveness* in line with e.g., Shortle et al. (2017, p. 3) to describe a queueing discipline's performance with regards to minimising waiting time (or some equivalent measure). However, other scholars on fair queueing has used *efficiency* (Avi-Itzhak and Levy 2004, p. 919) or *queue-efficiency* (Chun 2016, p. 10) to describe the same property.

Generally considered, there is no consensus in the fair queueing literature on the use and definition of efficiency. For instance, some authors such as Kayı and Ramaekers (2010) and Chun (2016, p. 10), who allow for compensatory transfers between customers, equate *efficiency* with *Pareto-efficiency*, and suggest that Pareto-efficiency can be decomposed into *queue-efficiency* (i.e. effectiveness, on my use) and *balancedness* (the sum of transfers between customers must sum to zero). Others, such as Tang, Yu and Li (2022, p. 1806), define (Pareto-)efficiency as the property that no customer can increase its resource allocation without harming at least one other user. I use *effectiveness* to describe performance with regards to waiting time and reserve the use of the term *efficiency* to describe the property of exhausting the resource-to-be-allocated, such as the capacity of a server. Hence, I shall refrain from using Pareto-efficiency altogether, and only use efficiency to mean good-exhaustion[7].

[6]Some scholars do the same (see e.g., Maccio, Hogg and Down (2018, p. 135)), whereas others try to satisfy both fairness and effectiveness simultaneously. Chun (2016, 11ff.) and Kayı and Ramaekers (2010) makes an attempt to the latter. They require monetary transfers between agents to offset the adverse effects on fairness from optimising effectiveness. The queueing systems which I treat do not allow for such transfers (see also Section 5.1).

[7]There is obviously a link between Pareto-efficiency and efficiency as good-exhaustion. For instance, on the definition of Tang, Yu and Li (2022), Pareto-efficiency implies my definition of efficiency as good-exhaustion: If the good is not exhausted, then the allocation cannot be Pareto-efficient, since it would be possible for a customer to increase their resource allocation

My definition of efficiency as good-exhaustion matches that of Wintein and Heilmann (2024a), whose definition of fairness I adopt. On this definition of fairness, efficiency — or good-exhaustion — is a fundamental requirement for fairness, as I explain in Section 4. In their papers, Wintein and Heilmann (2024a; 2024b) demonstrate a close kinship between the fairness literature of philosophy and the 'bankruptcy' literature of economics. Especially within the latter, the notion of efficiency, that I employ here, is engrained. For instance, Thomson (2003, p. 266) and Casas-Méndez, Fragnelli and García-Jurado (2011, p. 161) have incorporated efficiency directly into the definition of an allocation rule. Thus, a rule is *defined* such that the sum of all recommended allocations sum to the total estate, i.e. implying that it exhausts the good (see also Wintein and Heilmann 2024b, p. 18).

Alas, *effectiveness* refers to a queueing discipline's performance with regards to minimising (average) waiting time and *efficiency* refers to good-exhaustion. Furthermore, when I talk of the *performance* of a queueing discipline, I refer to both effectiveness *and* fairness[8]. Fairness, in turn, requires efficiency (understood as good-exhaustion).

3.3 When is fairness relevant for queueing?

Fairness and effectiveness are relevant concerns, whenever we care about doing good and whenever we care about spending fewer resources, respectively. But both concerns are not always equally important. For instance, concerns over fairness may be fully irrelevant when considering bottles in a filling line, but may be essential when dialytic patients are queueing for a new kidney.

This reflects the fact that many different types of entities can queue, but that fairness does not apply equally to them all[9]. Since fairness is a moral concept, it only applies to morally considerable beings. Thus, fairness in queueing is relevant if and only if the entities in the queue are morally considerable beings. This point is often overlooked by many scholars in the fair queueing literature.

This point caries a bit of terminology with it. I use the term 'client' to describe any entity — morally considerable or not — that is asking for a task to be completed by a server. It may thus be an actual client in a supermarket, it can be a job to be run on a computer or an empty bottle in a bottle filling line waiting to be filled. I use 'server' to describe the entity that serves the client. It may be a cashier at the supermarket, a literal server in a computer network or the bottle filling machine that fills bottles with soda. Finally, I use the term 'customer' to determine the subset of clients who are subject to moral considerations and thus fairness concerns. I leave it to the reader to determine exactly when some entity becomes subject to concerns regarding fairness[10]. The answer to this question

without harming other customers. The concepts are not identical, as it does not necessarily follow that an allocation that exhausts the good is Pareto-efficient.

[8]Decomposing performance into both effectiveness and fairness, as I do here, goes against the common conflation of performance with effectiveness seen in a large part of the literature on queueing (cf. e.g., Maccio, Hogg and Down (2018) and Wierman and Harchol-Balter (2003)). Avi-Itzhak, Brosh and Levy (2007, p. 1127) concur that equating performance with effectiveness is wrong.

[9]Thanks to Ruth Korte for enlightening comments on a previous draft of this thesis, and for pointing this out to me.

[10]One common option is to assume that *sentience* is required for something to be a morally considerable being (see for instance Kagan (2019)). This does not necessarily imply that

only affects the scope of my argument, but not its validity.

Queues are used for scheduling tasks and distributing service in most parts of society. Their applications range from assigning treatment to patients in hospitals to prioritising packages of communication on computer networks. The main share of the literature on queueing has been devoted to improving the effectiveness of queues. But optimising queues with respect to effectiveness may come at the cost of treating some customers unfairly. This aspect is often overlooked in the literature. Sometimes compromising fairness is acceptable, but often it is not[11]. In order to assess the fairness of a queueing scenario, we need a measure for assessing the fairness of queues. This is what I seek to provide here.

4 Fairness

Fairness is often used to mean different things. As stated in the previous section, this is particularly true for the literature on fair queueing, where various, ambiguous and imprecise definitions of fairness circulate. In order to develop a measure of fairness, a careful definition of fairness is needed. This is what I present in this section.

My definition of fairness is adopted from Wintein and Heilmann (2024a), who extend one of the most widely discussed accounts of fairness, provided by John Broome, to include an absolute dimension of fairness. Additionally, they provide the *Absolute Priority Rule* as an operationalisation of this definition. Beyond its broad appeal, there are three reasons for considering the Broomean account of fairness. First, it is concerned with fairness in and by itself as opposed to considering fairness as a means to theorise about something else. Second, it has a wide potential for practical application, as I shall demonstrate in this thesis. In particular its reliance on the notion of claims is useful in extending the account to the domain of queueing. And third, it integrates well with my view that queueing problems can be regarded as a special case of fair division problems.

4.1 A definition of fairness

Fairness is a subdivision of justice. Whereas justice is concerned with the elimination of arbitrary distinctions, fairness is concerned with the right dealings between persons who are cooperating or competing with each other (Rawls 1958, 164f.). Hence, one speaks of fair games, fair competitions and fair bargains. Questions of fairness often arise when free people without authority of

customers must by all means be sentient. I shall say that non-sentient things can be subject to moral considerations, if they stand in relation to a morally considered being in such a way that unfair treatment to the thing instantiates an unfairness to the being. For instance, concerns of fairness does not extend to an e-mail *per se* (because an e-mail is non-sentient), but I think it is sensible to talk of unfair treatment of an e-mail (in a queue, say) in its function of being *my* e-mail. To be clear, the unfairness is not done to the e-mail, it is done to me (by proxy) — but for brevity's sake, I shall allow for discussion of unfairness done to things. Just keep in mind that this implicitly assumes that the thing in question stands in a relevant relation to some morally considerable being, such as a human being.

[11] I leave it to the reader to assess the relative importance of effectiveness and fairness. I believe that one may consider both fairness and effectiveness as goods to be weighed against each other. Thanks to John Broome for helpful discussion on this point.

each other are engaging in some joint activity and seek to divide burdens and benefits between them.

Importantly, fairness is a relational moral concept. Consider the anecdote mentioned by Piller (2017, p. 233) of Sydney Morgenbesser, former professor of philosophy at Columbia University, who was beaten by the police when joining a human chain during student uprisings in 1968. Asked about how he felt, he should allegedly have said that it was unjust to beat innocent protesters, but not unfair as everybody got beaten. Thus, comparisons seem to be central to fairness, but not necessarily so to general justice.

Many theorists argue define fairness as a *purely* relational notion. However, I suggest that it makes sense to distinguish a comparative and an absolute dimension of fairness. This means that it is not only important how the good in question is allocated, but also *that* it is allocated. It is unfair not to allocate a good if one has the possibility to do so. I return to this point in Section 4.3.

4.1.1 Substantial fairness

There exists a minimal understanding of fairness as the consistent interpretation and application of rules. I call this *formal fairness*, in line with Hooker (2005)[12]. Formal fairness requires that rules are interpreted and applied equally to all individuals. Assume that there's a rule that says that whoever rolls the highest number with a die goes first, when playing a board game. Then, it is *formally unfair* if that person is not allowed to go first. In this way, queueing can be seen as an instantiation of formal fairness. A queueing discipline is essentially a rule that determines the order that people are served in. If the rule is not applied equally to all customers, then it is not *formally fair*.

But rules can be unfair. Suppose for instance that there's a rule that women should never be allowed to go first, when playing a board game. Because there's no morally relevant distinction between men and women, this rule is *substantially unfair*. This means that formal fairness can be satisfied — the rule might be applied consistently — without substantial fairness being satisfied. In other words, formal fairness can preclude substantial fairness depending on the rule.

Note that some scholars, such as Avi-Itzhak, Levy and Raz (2011), motivate their search for a fairness measure for queueing on the reason that queues provide 'fair service' to customers. Essentially, that queues serve to satisfy formal fairness. Spelling out the difference between formal and substantial fairness, as I do here, accentuates the difference between the formal fairness of queues and the substantial fairness of different queueing disciplines. It shows that the motivation for a measure of fairness for queueing must primarily rely on concerns of substantial fairness, rather than formal fairness.

As a result, I'm only concerned with substantial fairness in this thesis. As observed by Hooker (2005, p. 330), the importances of formal fairness fades significantly in the face of substantive unfairness. Thus, I'm not concerned with whether or not queueing disciplines are consistently applied — I simply assume that they are. Instead, I focus on whether the rules are substantially fair and to what degree.

[12]This could alternatively be called *procedural* fairness, in line with procedural justice (cf. Parfit (2000, p. 89)).

4.1.2 Fairness as an objective moral notion

Substantial fairness can be defined as either an objective or subjective moral notion. Taking fairness to be objective, as I do, means that fairness is not determined by a person's tastes, interests or preferences. Instead, it is supposed to be an impartial and objective evaluative criterion. In particular, the definition of fairness that I introduce below, pertains to objectivity by relying on individual 'claims' that are meant to be universally accepted and intersubjective. This understanding of fairness contrasts with *subjective fairness* that relies on individual preferences and is concerned with fairness from the individual's point of view. For instance, the 'no envy' definition of fairness, that requires that no individual in a division problem would prefer the share of another individual, is an example of such subjective fairness perspective (Wintein and Heilmann 2021).

In this paper, I treat fairness as an objective moral notion. I do so both because I'm interested in obtaining an objective criterion for assessing the fairness of queues, and because the implementation of a subjective notion can be problematic due to biases between individuals perceiving the same situation differently, or due to lack of consistency across different situations. This is not to say that objective fairness is problem-free. But the problems, I believe, are of a different and less intrusive character. I discuss some of these problems, such as the measurement of claims, in Section 4.3.1.

4.2 Broomean fairness

One of the most widely discussed and sophisticated notions of fairness is developed by John Broome and serves as the basis for the definition of fairness, I adopt. Broome's theory is related to the distribution of a good between people (Broome 1990, p. 87) and carries the central statement that *fairness requires that claims should be satisfied in proportion to their strength* (Broome 1990, p. 95). As such, this can be thought of as the generalisation of the core principle that equal claims should be treated equally (Piller 2017; Broome 1991, p. 196).

Claims are 'duties' owed to the individual. They are reasons for why a claimant should receive a good. Hence, Broome distinguishes claims from other reasons such as teleological reasons or side constraints. Teleology demands that a good is allocated such that it maximises overall benefits, and would thus weigh up one candidate's teleological reasons against those of another to determine the maximal benefit. Side constraints, on the other hand, determine what ought to be done with the good and therefore do not incur weighing of reasons (Kirkpatrick and Eastwood 2015, 83f.). Rights, for instance, are typically considered to be side constraints (Broome 1990, p. 91). Broome argues that claims are the appropriate reasons to consider when dealing with fairness concerns and that claims are owed to the *individual* rather than to the general good as some utilitarian perspective suggests (Broome 1984, 44f.). Broome does not provide a detailed account of what claims are, nor exactly what constitutes claims, but as Piller (2017, p. 216) argues we can understand these claims intuitively, or pre-theoretically.

The goods considered by Broome can be either divisible or indivisible. Whenever goods are divisible, fair allocation can typically be reached by dividing the good.

However, in the indivisible case it is not possible to divide the good to ensure a fair allocation. Imagine the allocation of a kidney between a group of dialytic patients: dividing the kidney would destroy it. A third of a kidney is of no use. To accommodate fairness, Broome suggests, one should treat claimants equally by giving them an equal chance (assuming they have equal claims) of receiving the kidney by choosing them randomly through an equal-weighted lottery. Although this does not ensure a fair allocation — some individuals receives a kidney and others (who have an equally large claim) do not — nobody was *treated* unfairly (Broome 1984, p. 45; Holm 2023, p. 271). As critics have pointed out, obtaining fairness in allocation of indivisible goods through lotteries has it caveats, however (see e.g. Kirkpatrick and Eastwood 2015). To avoid dealing with these caveats, I treat only goods that are divisible, in this paper.

Fairness depends on how one individual is treated *in comparison* to others. Particularly, Broome takes fairness to be *purely* relational. The claims, an individual holds, exists only in that particular context and in relation to the claims of the other candidates of that good, Broome (1984, p. 43) argues. And so, fairness does not necessarily require that claims are satisfied: proportional satisfaction of claims might involve no satisfaction at all (Holm 2023, p. 269). To have a fairness claim does not mean that the individual has a general right to the good. It only means that if the good to be allocated is made available for some people, every person with a fairness claim has the right to be among them (Broome 1984, p. 43). Broome however, is not silent on the satisfaction of claims: 'To be sure, all is not well if they get none at all. For each claimant there is at least one reason why she should have some of the good: the reason that constitutes her claim. Claims should be satisfied, therefore. But it is not *unfair* if they are not, provided everyone is treated proportionally.' (Broome 1990, p. 95). If there are more claimants to a kidney, it is *morally* unacceptable not to distribute the kidney, but it is not *unfair* not to distribute the kidney.

This difference between moral and fair demonstrates how Broome suggests that fairness is only one reason that counts towards determining the right allocation. As such, Broome's theory of fairness is not (and is neither meant to be) a complete moral theory. Fairness only determines one aspect of how one ought to act (Piller 2017, p. 215). There might be other reasons for choosing a particular allocation; the fair allocation need not be the *right* allocation. This point is important to keep in mind throughout the remainder of this paper: I only provide an account of what constitutes *fair* queueing, not *just* queueing.

4.3 How to be absolutely fair

Whereas Broome raises the concern for satisfying claims outside his account of fairness, Wintein and Heilmann (2024a; 2024b) suggest that the concern for satisfaction of claims can be regarded as a matter of fairness in itself. On their definition, fairness also possesses an absolute dimension beyond the strictly comparative dimension, and hence good-exhaustion is also a concern for fairness, they argue. If a good is not distributed, this is not only a problem for justice, it is also *unfair* that the good is not distributed. Consider the logic of Hooker (2005, 339f.) by the following example: Imagine that I owe 100 to Susanna and 100 to Figaro. I promised to repay my debts today, and with me I have enough to repay both of them their full claims. However, rather than repaying them, I decide to keep the money. Have I treated both Susanna and Figaro *equally*?

Sure. But I have not treated them *fairly.* Going forward, I define fairness such that it encompasses an absolute dimension. Despite not adhering to the purely comparative account of fairness that Broome subscribes to, I shall still from place to place refer to it as 'Broomean fairness' (or simply 'fairness') due to its reliance on the proportional satisfaction of claims.

4.3.1 Claims

In order to examine the implications of adopting an absolute notion of fairness, Wintein and Heilmann (2024a) clarify the notion of a claim. In particular, they argue that in addition to its strength, a claim has an *amount.* The amount determines the value required for full satisfaction of the claim, and thus also constitutes the constraint of the claim. Whenever the claimant receives a share of the good equal to the amount of the claim, we say that the claim has been *fully* satisfied. Any receivings beyond the amount of the claim do not serve towards the satisfaction of the claim (but may serve towards satisfying other things, such as the claimant's preferences) (Wintein and Heilmann 2024a, 7f.). Although claims' amounts are not explicitly mentioned by Broome, this is implicit from the notion of *satisfaction* which is explicitly stated by Broome. Amounts of claims are needed to determine the degree of satisfaction of the claim (Wintein and Heilmann 2024a, 4f.).

Moreover, claims have *strengths.* The strength of claims measures how strong a reason is for satisfying a particular claim and is a purely comparative notion. Suppose two individuals, Don G. and Leporello, work towards a common task that produces a particular outcome, say, of 1,003. Both individuals have a claim of the same amount to the outcome of 1,003. Suppose now that Leporello spent nine hours towards completing the shared task, and Don spent only one. Then the reasons for satisfying Leporello are stronger *compared to* the reasons of Don. Specifically (other things equal), it is nine times stronger (Wintein and Heilmann 2024a, p. 5; Piller 2017, p. 234).

In sum, Wintein and Heilmann (2024a, p. 5) define a claim as consisting of both an *amount* and a *strength.* Moreover, they divide claims into *notional* and *absolute* claims. Absolute claims do not depend on the comparative satisfaction of other individuals claims: Absolute claims require full satisfaction, regardless of the satisfaction of other individuals' claims. Suppose you owe 91 to Don and 31 to Leporello. Don's claim to 91 is absolute and thus independent of any repayment you may make to Leporello. It is non-comparatively unfair to not repay Don.

Notional claims, on the other hand, do not require full satisfaction necessarily. Their proper satisfaction depends on their comparative satisfaction. Don and Leporello both worked towards obtaining the outcome of 1,003. As long as Leporello receives a share of the outcome that is nine times larger (because he spent nine hours on the task, and Don only one) than that of Don, all is good; he has received his 'fair share'.

4.3.2 What reasons constitute claims?

It is widely debated, what reasons constitute claims. Broome suggests that if we accept *desert* as a moral reason, then we might also accept it as a claim. However, as Hooker (2005, p. 343) points out, we often do not deal with desert in

a proportional way. Rather, we suggest that the most deserving candidate gets all the good. The main prize in a competition is awarded to the best candidate, it is not distributed proportionally to desert.

Similarly, *need* is dubious as a constituent of claims. Hooker (2005, 341ff.) argues that satisfying needs is more often than not a proxy for benefitting the worst off. But benefitting the worst off is not a matter of fairness, as Broome argues. If needs, in one way or the other, are not used as proxy for the worst of, then it seems that needs do not really constitute morally important reasons.

Ultimately, Broome (1990, p. 93) admits that it is not so clear what particular reasons are claims, and which are not. As Piller (2017, 218f.) notes, Broome's theory is incomplete with regards to providing proper definitions of what claims are, and what they are claims to. Therefore, Broome's theory is quite stunted when applied in practice, as there may be multiple candidates for claims depending on definitions. Imagine that the young Pamina and the old Sarastro are held hostage, but can be saved by a handsome prince. The prince however can only save one of them and wants to make his decision fairly. He therefore tries to determine each person's claim to be saved from hostage. If individuals have a claim to a long life in freedom, then the prince should save the young Pamina rather than the old Sarastro, who has already had a long life. If, on the other hand, all people have an equal claim to be saved from hostage then the age difference between Sarastro and Pamina does not matter.

I do not have any hopes on settling the general controversy of what reasons constitute claims here. In Section 5, I shall provide suggestions to possible determinants of claims in the specific domain of queueing, but will eventually just assume that claims can be determined.

4.3.3 Can we measure claims?

Another issue with claims is their measurement. Kirkpatrick and Eastwood (2015) have argued that it is almost impossible to measure claim strength accurately[13]. Piller (2017, 234 ff.) is less sceptical. First, Piller (2017) argues that it is sometimes rather clear how to determine the strength of claims, such as when one individual works nine hours and another works one hour towards a shared task. In other situations, it is less clear. Piller proposes two possible avenues: We can give up the requirement for having *exact* values of strengths and apply inaccurate ways to determine strengths. Or we can allow for interpersonal comparisons of utility and apply expected utility theory to determine strengths of claims. In case neither option is found appealing, Piller (2017, p. 236) argues that we may accept Broome's theory as incomplete on the matter, however this does not undermine the theory, Piller says.

Kirkpatrick and Eastwood (2015, 89f.) argue that a possible solution to their objection regarding measurement of claim strength is to use authorities or rules. This solution is however ruled out by Broome himself. In fact it is part of the motivation, Broome provides for his account, that authorities and rules have the potential to be unfair. However, Kirkpatrick and Eastwood (2015, p. 90) argue

[13]To be precise, Kirkpatrick and Eastwood (2015, 86f.) argue that it is not possible to calculate the appropriate *weights* for a lottery between claimants to an indivisible good. Fairness requires, Broome suggest, that the *weights* of a lottery should be determined by the *strength* of claims. I see no reason why the arguments of Kirkpatrick and Eastwood (2015) should not apply to both weights and strengths.

that such use of authorities is needed to formulate the theory itself. Kirkpatrick and Eastwood (2015, p. 90) take this to imply that Broome thus undermines his own theory. Someone less sceptical might suggest that it at least compromises the use of experts for determining the actual values of strengths of claims.

Once more, I shall not dwell to much with this controversy. In the remainder of this project, I shall work out the implication *assuming* that both claims amounts and their strengths can be measured.

4.4 The Absolute Priority Rule

With the addition of an absolute dimension to fairness, Wintein and Heilmann (2024a) provide a *Fairness Formula* as an update to the Broomean requirement of proportionality. According to the formula, fairness requires that:

Good-exhaustion (or efficiency) (I) absolute claims should be satisfied to the greatest possible extent as long as no one receives more than they have a claim to,

Proportionality (II) individual claims (both notional and absolute) should be satisfied in proportion to their strength, and

Conflict-resolution (III) the first requirement should be satisfied first whenever it conflicts with the second, while respecting the second requirement to the greatest extent possible.

Wintein and Heilmann (2024a, 15ff.) then propose a rule that operationalises the Fairness Formula. The *Absolute Priority Rule* is defined as follows: Given the estate of a good (for instance, server *C*apacity) C and a set of customers $N = \{1, ..., n\}$. Define the truncated estate as $\bar{C} = \min\{C, \sum_{j \in N} a_j\}$, i.e. the minimum of the total estate and the sum of all claims' amounts. Let C_R denote the remaining estate to be allocated among the set N_R remaining individuals.

The fair allocation, as recommended by the absolute priority rule, is determined as follows:

1. Set initially $C_R = \bar{C}$ and $N_R = N$.
2. Compute the allocation $x_i = \frac{s_i \cdot a_i}{\sum_{j \in N_R} s_j \cdot a_j} \cdot C_R$, where s_i denotes the strength of customer i's claim and a_i denotes the amount.
3. If there exists no customer for which the allocation $x_i > a_i$ exceeds the customers claim, then allot x_i for each i in N_R. Otherwise, proceed to step (4).
4. For all customers for which the allocation $x_i > a_i$ exceeds the customers claim, allot the individual a_i. Update the remaining server capacity C_R and the remaining number of customers N_R accordingly. Repeat steps (2) and (3).

The absolute priority rule becomes a cornerstone in the development of the Broomean Fairness Measure for Queueing. But before introducing the measure, I need to first explain how fairness can be conceptualised for queueing. This is what I turn to now.

5 Fairness in Queueing

In this section, I show how fairness, as defined in the previous section, can be applied in queueing. I do so in two steps. First, I argue that certain queueing problems can formally be represented as what I call a 'queueing structure' that represents the relevant features of the queueing problem. Second, I argue that one can map these queueing structures to a 'Broomean problem'. A Broomean problem is a *fairness structure* in which the relevant characteristics of the target system are interpreted in terms of theoretical fairness notation (thus, introducing terms such as 'claims', 'amounts' and 'strengths of claims'). As such, the queueing problems, that can be represented as queueing structures, can be regarded as *problems of fair division.*

In other words, I am providing a representation that extracts the important characteristics of queueing understood as a problem of fair division. This representation serves as the basis for applying a *fairness function.* Hence, a third step is to provide of a fairness function. A fairness function is a rule that provides an allocation of the resource for each of the fairness structures within its domain (Wintein and Heilmann 2024a, 10f.). Given the representation of the queueing problem as a Broomean problem, it suggested by Wintein and Heilmann (2024a) that the appropriate fairness function is the *absolute priority rule.*

5.1 Queueing problems

In Example 2.1, I presented a simple queueing problem of four mathematicians who needed computing power to solve their individual mathematical problems. Each mathematician had different needs for computing power and the computer was not restricted to serving one mathematician at a time, but could serve all of them at once. In the example, I argued that the information relevant to the Resource Allocation Queueing Fairness Measure (RAQFM) of Raz, Levy and Avi-Itzhak (2004) for the first epoch could be stated as:

$$\mathcal{Q}_1^{\mathrm{FM}} = \langle 1000, \{A, B, C, D\}, (1/4, 1/4, 1/4, 1/4)\rangle_1 \,, \tag{2.1}$$

where $(1/4, 1/4, 1/4, 1/4)$ indicates that there are no reasons for assigning any one mathematician more of the total computer's capacity than the others, as assumed by Raz, Levy and Avi-Itzhak (2004).

Avi-Itzhak, Levy and Raz (2008) argue that queueing problems can be interpreted as the allocation of a resource amongst a number of customers, considered in a given epoch t. The formulation given in Equation (2.1) makes this interpretation explicit. I argue, however, that the RAQFM disregards relevant information by not taking into consideration the individual requirements for service. Therefore, I suggest that a better representation of the queueing problem in the first epoch is this:

$$\mathcal{Q}_1^{\mathrm{FM}\star} = \langle 1000, \{A, B, C, D\}, (600, 450, 300, 150)\,, (1/4, 1/4, 1/4, 1/4)\rangle_1 \,, \tag{5.1}$$

where the individual service requirements are included.

Going forward, I refer to such a representation as a 'queueing structure'. A queueing structure records the information from the queueing problem that

is relevant for determining the fair allocation of the server's capacity. Before discussing how queueing problems should be represented in detail, let me first provide another — less stylised — example of a queueing problem.

Example 5.1. Queueing on packet switch systems In a seminal paper on fair queueing, Nagle (1987) describes a packet switch system with infinite storage. Packet switch systems are commonly used for data transferring in telecommunications, where data is grouped into packets in order to be transmitted via a digital network. The method is amongst other things used for data transferring on the internet and also in modern mobile phone technologies such as GSM and LTE.

A switch, or gateway, has incoming and outcoming links. Packets of data are sent by sources, and arrive on the incoming link on the switch. Every packet is assigned an outcoming link, which serves packets at a fixed data rate.

The queueing problem, considered by Nagle (1987), consists of determining the optimal queueing discipline for serving the data packets on the switch. But notably, when determining the optimal discipline, Nagle (1987) not only considers the effectiveness (the extent of congestion on the system), but also the fairness of the queueing discipline. The solution offered by Nagle (1987) has become known as the round-robin discipline. But as Demers, Keshav and Shenker (1989) note in a response to Nagle (1987), this solution relies on the assumption that data packets are of the same size. In practice, they are often not.

Demers, Keshav and Shenker (1989) thus provide an adaptation of Nagle's initial model with unequal packet sizes. Specifically, Demers, Keshav and Shenker (1989, Scenario 1) examine a set of different queueing scenarios. In the first scenario, there are four clients — two Telnet sources (T_1 and T_2) and two FTPs (F_1 and F_2) — on a network with a single server (S). In a given time frame of five second, the Telnet sources each send one 40 byte packet and every FTPs send 17 packets of 1000 bytes[14]. The server sits behind a gateway (GW) that can process 56 kilobytes per second, or 208 kilobytes per every five seconds. According to Demers, Keshav and Shenker (1989), each source has an equal 'right' to the servers capacity[15]. Figure 5.1 provides a (slightly modified) version figure from Demers, Keshav and Shenker (1989, p. 7).

Like Raz, Levy and Avi-Itzhak (2004), both Demers, Keshav and Shenker (1989) and Nagle (1987) are assuming that queueing can be interpreted as the allocation of a scarce resource. The nature of the packet switch queueing problem (PS) allows us to represent it for the first epoch by the following queueing structure:

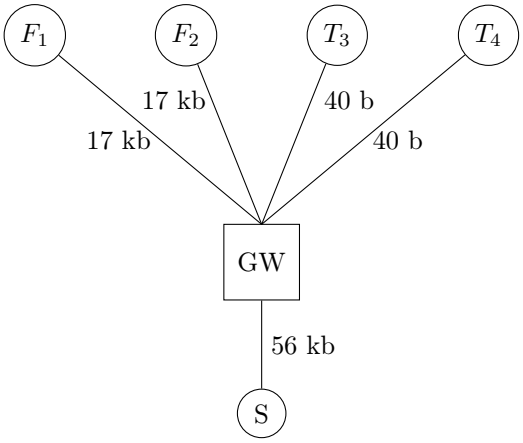
$$\mathcal{Q}_1^{\mathrm{PS}} = \langle 208000, \{T_1, T_2, F_1, F_2\}, (40, 40, 17000, 17000), (1/4, 1/4, 1/4, 1/4)\rangle_1 .$$

Obviously, not all queues can be represented as queueing structures, nor benefit from attempts to do so. For instance, fair queueing has also been addressed as a question of social choice. The queueing problem considered by

[14]To be accurate, the number of packets sent by the FPTs is controlled by a flow control algorithm. For simplicity, I'm simply taking the average number of packets per five second. In their simulation, 1746 packets are sent in the course of 500 seconds, which means that on average $17.46 \approx 17$ packets are sent every five seconds. I refer to Demers, Keshav and Shenker (1989, p. 7) for details.

[15]To both Demers, Keshav and Shenker (1989) and Nagle (1987) *fair* is the same as *equal*. As I have already argued, this is not warranted.

Figure 5.1: Diagram of the considered queueing system



Notes: Graphical representation of the queueing scenario described in Demers, Keshav and Shenker (1989, Scenario 1), with slight modifications. There are four packet sources that have different requirements for service as well as one server that sits behind a gateway with a capacity of 208 kbytes per every five seconds. 1 kb = 1000 b(yte).

social choice theorists is typically quite different, however. In general terms, the queueing problem in social choice theory such as described by Chun (2016) or Kayı and Ramaekers (2010) considers a fixed number of agents in a facility in which only one customer can be served at a time. Agents incur waiting time costs, thus being served at a later stage has a higher cost to the agent's utility than being served earlier. Notably, this formulation of the queueing problem allows for monetary transfers between agents to compensate for differences in waiting costs. The queueing structure, I have described, does not allow for such monetary transfers and hence these queueing problems are not fit for representation. This does not concern me too much as I believe it is a rare characteristic of queueing systems to have monetary transfers implemented. Proceeding, I will only consider queueing problems that can be represented as queueing structures, like $\mathcal{Q}^{PS}$ or $\mathcal{Q}^{FM}$.

5.2 Queueing as a fair division problem

Fair division problems are problems concerned with the question of how one should divide a scarce good, in order to be fair (Wintein and Heilmann 2020). In Section 4, I introduced a claims approach to dealing with such problems. Wintein and Heilmann (2024a, p. 10), whose work I rely on, suggest that to solve a fair division problem — that is to determine which allocation of the resource is *fair* — one must represent the problem as a *fairness structure* to which a *fairness function* can be applied. The fairness function assigns a fair allocation for all the structures within its domain. In light of my definition of fairness, the structures I consider is 'Broomean problems' (Wintein and Heilmann 2024a, p. 11). Put formally, a Broomean problem can be written as the tuple

$$\mathcal{B} = \langle C, N, a, s \rangle,$$

where $C > 0$ is the resource to be divided between a set N of claimants where each claimant $i \in N$ has a claim with an amount $a_i > 0$ and strength $s_i > 0$.

The alert reader should see immediately that the representation of fair division problems as Broomean problems is very similar to the representation of queueing problems as queueing structures. This is no coincidence. In fact, I'm suggesting that any queueing problem that can be represented as a queueing structure can be mapped as a Broomean problem. This is possible whenever queueing can be interpreted as a problem of allocating a scarce resource, i.e. a fair division problem.

By interpreting queueing as a problem of resource allocation, authors such as Raz, Levy and Avi-Itzhak (2004), Demers, Keshav and Shenker (1989) and Nagle (1987) are thus implicitly taking the steps that I make explicit here. (Remember that the fairness measure of Raz, Levy and Avi-Itzhak (2004) is called a '*Resource Allocation* Queueing Fairness Measures')[16]. The next step that I take, but previous authors have not, is to employ the literature on solving fair division problems for the purposes of queueing.

But before doing so, I must first account for how queueing problems are mapped to Broomean problems. Because of the one-to-one mapping between queueing structures and Broomean problems, I shall say skip the intermediary step of representing queueing problems as queueing structures. Going forward, I discuss how queueing problems can be represented directly as Broomean problems.

Broomean problems comprises four elements. Accounting for the first two elements, C and N, is straightforward. The capacity of a server (e.g., the available computing power in a given time span) can be considered a resource to be divided between a set of customers waiting in line. However, accounting for customers' claim amount and claim strength is somewhat more intricate. This is what I turn to now.

5.3 What constitutes claims in queueing?

I've argued so far that claims are reasons for assigning a resource to a claimant. In the domain of queueing, I suggest that the mere *requirement for service* provides a reason for assigning resources (or service[17]) to a customer, and therefore constitutes a claim. Since claims are 'duties owed to the individual', we may say that service is a duty *owed* by the server to the customer.

Suppose you have to send a packet of data via a switch. This is a requirement[18] for service that you make to the server on the switch. According to my suggestion, you thus have a claim to be served by the server. Critics may argue that simply requiring service does not mean that it is owed to you. The server in a restaurant, say, does not *owe* the customer service. She has no duty to provide service to her customers. I do not agree. When entering a restaurant and ordering a meal, you enter into an exchange with the establishment in which you provide *a promise to pay* for your meal and the service provided.

[16]For comparison, see also Appendix A.1 for alternative interpretations of queueing.

[17]I refer to *service* broadly as some sort of work being done for someone (or something) else.

[18]To avoid confusion, I avoid using the term 'need' for its close connotation with maximisation of benefits (see e.g. Hooker 2005, p. 342) and for the ambiguity in regards to whether it is related to claims' amounts or strengths.

The restaurant in turn provides you food and service. The promise of payment is enough, I believe, to require service and for the server to owe service to you. Obviously, it also implies that the server has a claim against you, namely the payment. Thus exchanges incurs duties on both parties. I do not consider the mere timing of the resolution of payment and service as important towards the establishment of duties.

However, claims can also be decoupled from any direct payment. Although they may often be linked, it is not necessary that claims are correlated with the amount that one has paid or promised to pay. Providing a promise to pay for service is sufficient but not necessary for a customer to have a claim against the server. Imagine a public hospital where no direct payment is necessary. The patient still has a claim to service from the doctor. Or imagine for instance being in a public library for which you have not paid to be a member (at least not directly) and asking the staff for help to find a book. If service requirement constitutes a claim, as I suggest, it implies that you have a claim to service, and your claim has an amount that is equal to the resources the staff need to spend in order to find the book. Some might find this counterintuitive, especially when considering very large service requirements: does it mean that if I need service to find a thousand books then I also have a claim to that service? In principle, yes. If the library is committed to being a library and committed to providing their users with help to find literature, then you as a user (or customer) have a claim to that service. However, and as I shall comment on below, the strength of the claim to help for finding a thousand books might by rather minute in comparison to other individuals' claims.

I thus take it to be definitory for a server to have a duty to serve, if the server wants to remain a server: It is the job of a *server* to *serve* and if the server does not serve, it is no longer adequate to call it a server. So to remain a server, we may say that the server has a *duty* to serve, or that the server *owes* service to the customer. Given that a claim is a 'duty owed to the individual', I argue that service requirement can indeed constitute a claim.

Importantly, the server must serve someone or something requiring service. Serving someone with no requirement of service is not service: A hairdresser offering to cut the hair of a bald person is not really offering service. This is regardless of whether the offer was given with the intention of providing service. Service requires that there must be some degree of acceptance of the service by the recipient. Finally, service requirement constitutes an absolute rather than a notional claim (cf. Section 4.3.1). This implies that the customer requires full satisfaction of the claim and that the satisfaction is not dependent on its relative satisfaction compared to other customers' claims.

5.4 What determines the amount of claims in queueing?

In Section 4.3.1, I mentioned that a claim consists of both an *amount* and a *strength*. If the requirement for service constitutes a claim, then the *amount* of the claim is equal to the resources required to fulfil the customers service requirement. If it takes 450 units of computing power to satisfy your requirement for service, then your claim has an amount of 450. That is your claim has an amount equal to the resources it requires to e.g., send the packets of data.

It often happens that customers have service requirements that vary in magnitude. In the case of a hospital, not all patients require the same amount of

treatment. This means that the amounts of claims will be different amongst customers, too. Disregarding claim strength for a moment, there are thus reasons for assigning a larger share of the resource to some individuals than to others. This contrasts the common assumption in the fair queueing literature that there are no particular reasons for assigning one individual more of a resource than another. For instance, Demers, Keshav and Shenker (1989, 3, their emphasis) assume that customers have 'equal *rights* to the resource' and Raz, Avi-Itzhak and Levy (2006, 1, my emphasis) argue that all customers '*deserve* equal service'.

It follows from proportionality (cf. Section 4.4) that equal claims (i.e., claims with equal amounts and equal strength) must be treated equally. Thus, the conclusions of Demers, Keshav and Shenker (1989) and Raz, Avi-Itzhak and Levy (2006) are congruent with the account of fairness I treat here, *insofar* as claims are equal. However, the assumption that all users have equal claims is unfounded, I believe. Despite being equal in all other regards, customers who differ in their service requirements are not equal. Unless it can be argued that service requirement is morally irrelevant, I do not believe it is appropriate to assume that all customers have equal claims to the resource. And because the individual service requirement is morally relevant for assessing the fair division of the server's capacity in a queueing problem, it should be represented by the fairness structure.

Worth noting, Demers, Keshav and Shenker (1989, p. 3) themselves argue that some users 'deserve' more bandwidth than others[19] — however they believe it is impossible to adequately establish such varying desert and therefore stick with the assumption of equal claims. When service requirement is constitutive of claims, this is no longer an issue.

5.5 What determines the strength of claims in queueing?

I turn now to the strength of claims and propose two features that determine claim strength. I do not make any illusions that these two features exhaust all possible determinants of claim strength. But I believe them to be important determinants. I will base the the succeeding work on the assumption that these two features together determine claim strength.

Seniority First, I suggest that *seniority* is a determinant of the strength of a claim. Seniority refers to the relative position in a queue as measured by rank. The higher the seniority, the lower the rank. If there are four customers A, B, C and D in a queue, and customer A arrived first and customer D last, then A has rank 1 and D has rank 4. This means that seniority is a comparative notion. Regardless of the exact time I've waited in the queue, if I arrived before you, my seniority is higher than yours and my claim will be stronger than yours (other things equal). Seniority implies that for two claims with equal amounts the claim with the higher seniority has the stronger claim to the resource.

Consider a scenario where two bachelors, Ferrando and Guglielmo, arrive in a queue to a public bathroom. Ferrando arrives shortly before Guglielmo. Assuming other things equal and that only one individual can use the bathroom

[19] Demers, Keshav and Shenker (1989) seem to use 'desert' interchangeably with 'rights'. This is another example conflating words with different moral meanings.

at a time, I believe most people would argue that Ferrando has the stronger claim to service, i.e. accessing the bathroom. Because of his higher seniority, Ferrando has a stronger claim to accessing the bathroom.

Remember, that Raz, Levy and Avi-Itzhak (2004) defined two 'principles of fairness' that I referenced in Section 2.2.1, suggesting that it is fair that jobs are served according to arrival times (seniority preference principle), and it is fair for large jobs to have large response times in comparison to short jobs (service requirement principle). It now shows that, although I agree with Avi-Itzhak and Levy (2004) that seniority is relevant to the fairness of queues, my implementation is much different from theirs. I argue that there may be other concerns than seniority, such as urgency, that may influence the strength of a claim and ultimately which allocations are considered fair and which are not. The *seniority preference principle* thus needs further qualification to work.

Urgency The second determinant of the strength of claims, I propose, is *urgency* of claim satisfaction. It refers to the objective importance[20] of claim satisfaction and typically depends on the realisation of some event with a positive or negative consequence. Urgency depends on the time until realisation of the event and the relative size of the consequence of the event. The shorter the time until the resolution of the event, and the greater the consequence of the event, the higher the urgency. Like seniority, urgency is also a comparative notion. The urgency of my service requirement is relative to yours.

Imagine once again that Ferrando and Guglielmo are queueing to a public bathroom. In this case, both individuals face the event of involuntary urination with the same negative consequence: having wet and smelly trousers. Both arrive at the same time, but getting to the bathroom is much more urgent to Guglielmo than it is to Ferrando. Again, I argue that Guglielmo has the stronger claim to the using the bathroom, because of the greater urgency connected to the satisfaction of his claim.

I use the term 'urgency' rather than 'need' to avoid any confusion of terms. Past users of the word have referred to needs as a general potential reason for claims (e.g. Broome (1990) and Hooker (2005)), or only as a determinant of claim strength, such as Wintein and Heilmann (2024a, p. 14). Wintein and Heilmann (2024a) consider an example where a person called Abram *'needs'* owed money twice as strongly as two other individuals and therefore has a *stronger* claim to it. I believe my use of urgency can substitute the use of 'needs' in Wintein and Heilmann (2024a) example: Getting the owed money is twice as urgent to Abram (maybe because he has a payment due, and not meeting the payment date has negative consequences for him) and therefore his claim is twice as strong.

I mentioned above that even very large service requirements could constitute claims, such as in the case where one individual requires help to find a thousand books at the public library. The adoption of urgency as a determinant of strength of claims may serve to avert scepticism towards such cases: it is quite likely that finding book number 501 is not very urgent in comparison to

[20]Scanlon (1975) uses a similar notion of urgency (or objective importance) in relation to moral judgments to account for the fact that some individual's claims to social resources are more urgent than others. Although the scope of Scanlon's project is more general — and thus his conclusions stronger — I take his argument as lending support to the idea of urgency as a moral consideration. See also Broome (1991, footnote 19).

the service requirement of other customers. Despite having a claim of a large amount, the strength of the claim is not necessarily very strong.

5.6 Measuring claims in queueing

As mentioned in Section 4.3.3, it is debated how the exact value of claims' amounts and their strengths are assessed in general. I believe that neither in the domain of queueing does there exist a clear answer to this question.

With regards to the amount of claims, applications in queueing theory usually takes the service requirement of a customer as given. Indeed, there are many cases where the service requirement is easy to assess. Suppose for instance that a source needs to send 1000 Gbit via a switch that can handle 1 Gbit per second. In this case, it is easy to determine the amount of the service requirement. In other cases, however, it is less clear and one must rely on past experiences. For instance, the exact processing time needed by a cashier in a supermarket is not previously determinable, but good approximations can be attained.

Seniority can be measured either as the relative position in a queue, or as the relative time spent queueing. Here, I argue that relative position, i.e. rank, should be used, so as not to double count time, when considering both urgency and seniority[21]. Suppose that we measured strength in the following way:

$$\text{strength} = w_1 \times \text{urgency (time} \times \text{consequence)} + w_2 \times \text{seniority (rank)},$$

where w_1 and w_2 are arbitrary weights[22]. Suppose seniority was also some derivative of time. In this case, both urgency and seniority would depend on time. But their dependency would be inversely correlated and might in some scenarios cancel out. Using rank ensures that all information morally relevant to strength is included in the measure of strength.

Note also that I assume away the possibility of misrepresentation of service requirements by customers, or lying. Maybe Guglielmo simply can't be bothered to wait in line at the bathroom and therefore pretends that getting access to the bathroom is much more urgent than it actually is. Here, I assume that claims are *objective* and can be obtained impartially. I therefore neglect this issue. Moreover, I believe the extent of the problem is dependent on the domain regarded. For instance, it seems unlikely that customers on a digital network exaggerates their need for service, whereas human beings might have obvious incentives to misrepresent their claims. The objectivity of claims implies that claims are made (externally) known as soon as they become (internally) known to the customer: Guglielmo will make his claim known as soon as it arises. He does not speculate in waiting, thus incurring greater urgency of his claim to the bathroom, before he makes his claim known.

[21] Thanks to Gideon Frey for pointing this out to me.

[22] In this thesis, I also disregard any concerns regarding the assessment of values of weights. We would also commonly assume some sort of transformation function on both urgency and seniority, i.e., $g(\text{urgency})$, where $g(\cdot)$ is some real-valued function. I disregard it to avoid convoluted expressions. I also disregard any other factors ε, other than urgency and seniority, affecting strength.

6 Fairness in Queueing: Applications

In the previous section, I argued that some queueing problems could be regarded as fair division problems and in turn represented as a Broomean problem. Wintein and Heilmann (2024a) have argued that the appropriate fairness function to apply to Broomean problems is the *absolute priority rule.* In this section, I provide a few examples of how one may do so. In all examples, I treat cases where there is initially not enough resources to serve all claimants in the queue immediately. There is $c = 1$ server capable of serving all customers at the same time. There are $n = 4$ individuals in the queue, and in all examples I assume that the server has a capacity of $E = 1000\,{}^{\mathrm{GBit}}/_{\mathrm{s}}$.

6.1 Example A: Claims with different amounts of equal strength.

Let me revisit the example of the four mathematicians, that I introduced in Section 2.2. The service requirement of the four mathematicians can be represented by the vector $(600, 450, 300, 150)$ and I assume that they all have the same seniority (they arrived in the queue at the same time) and their urgency is the same. This means their claims all have the same strength. As claim strengths are strictly comparative it shall not affect results if I normalise them such that $\sum_{i \in N} s_i = 1$. In this case, $\forall i \in N : s_i = {}^1/_4$. This implies that we can now represent the problem as the Broomean problem:

$$\mathcal{B}_1^A = \left\langle 1000, \{1,2,3,4\}, (600, 450, 300, 150), \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \right\rangle$$

As there is no mathematician whose allocation exceeds their claim ($\nexists i : P(\mathcal{B}_1^A)_i > a_i$), the allocation suggested by the *absolute priority rule* $P^\dagger$ (cf. Section 4.4) coincides with the *weighted proportional rule* P given as:

$$P(\mathcal{B}_1^A)_i = \frac{s_i \cdot a_i}{\sum_{j \in N_R} s_j \cdot a_j} \cdot C_R.$$

Table 6.1a provides an overview of the customers' initial claims, their strength and the allocation of resources by the absolute priority rule.

Table 6.1: Example A

(a) Initial allocation

i	Claims a_i	Strength s_i	Allocation by $P^\dagger$
1	600	1/4	400
2	450	1/4	300
3	300	1/4	200
4	150	1/4	100

(b) After one second

i	Claims a_i	Strength s_i	Allocation by $P^\dagger$
1	200	1/4	200
2	150	1/4	150
3	100	1/4	100
4	50	1/4	50

Notes: The first panel shows the initial claims and allocation, the second panel shows the claims and suggested allocation after one second of service. In this example, claims differ but the strength of all customers' claims are equal. The allocation of resources (per second) coincides between the weighted proportional rule and the absolute priority rule.

After a second of service (remember the server has a capacity of 1000 GBit per second), the mathematicians' claims will be updated. To obtain the new claims I subtract the allocated resources in the first second from the initial service requirement. Essentially, one is then faced with a new fair division problem, representable by the Broomean problem

$$\mathcal{B}_2^A = \left\langle 1000, \{1,2,3,4\}, (200, 150, 100, 50), \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \right\rangle,$$

which is also represented in Table 6.1b. At this stage, there is enough capacity to serve all of the mathematicians immediately, so every mathematician is assigned service equal to their service requirement.

6.2 Example B: Claims with same amounts but different strengths

Consider now the scenario where the urgency of one claim is much larger than the others. Suppose for instance that you're making a video call and sending off three e-mails (of the same size) at the same time via the same network switch. It is much more important that there is no latency on the video call compared to a bit of latency on the e-mails. We may represent the urgency of the video call by assigning it a strength of 7/10 and the other three mails a strength of 1/10. Assume that the service requirements are the same as before, such that the queueing problem can be represented by Table 6.2a or by the tuple:

$$\mathcal{B}_1^B = \left\langle 1000, \{1,2,3,4\}, (400, 400, 400, 400), \left(\frac{7}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right) \right\rangle$$

Table 6.2: Example B

(a) Initial allocation

i	Claims a_i	Strength s_i	Allocation by P	Allocation by $P^\dagger$
1	400	7/10	549.0	400
2	400	1/10	78.4	200
3	400	1/10	78.4	200
4	400	1/10	78.4	200

(b) After one second

i	Claims a_i	Strength s_i	Allocation by P	Allocation by $P^\dagger$
2	200	1/3	333.3	200
3	200	1/3	333.3	200
4	200	1/3	333.3	200

Notes: In this example, one customer has a claim with much higher strength than the other customers. Here, the weighted proportional rule assigns more of the resource to the first customer than the customer requires. The absolute priority rule thus assigns the first customer the full service requirement and then recalculates the amount of the resource to be allocated.

In this case, the weighted proportionality rule will assign customer $i = 1$ a proportion of the total bandwidth that exceeds its claim. The absolute priority rule requires that we assign the first customer her full claim (but no more), update the remaining capacity and repeat the calculation according to the weighted proportional rule. The final allocation can be seen from the last column of Table 6.2.

The first customer has its full claim satisfied. Assuming that no new customers arrive, nor make additional claims, we face a new problem given as represented in Table 6.2b or as represented by the Broomean problem

$$\mathcal{B}_2^B = \left\langle 1000, \{2,3,4\}, (200,200,200), \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \right\rangle.$$

We see that the strength of the claims (and the number of customers queueing) have now adapted in light of the customer who left after having its claims satisfied. In this case, there's enough of the resource to satisfy all remaining claims.

6.3 Example C: A dynamic queue

In this last example, I do not fix the total number of customers in the population and further assume that the arrival times of the customers differ. Specifically, the customers arrive at one second intervals. Imagine that it takes 1350 Gbit to satisfy the service requirement of a customer. Hence, all customers have claims of 1350, and their claims are equally urgent such that it is only their seniority that differs. In the first second, there is only one customer present who is assigned the full capacity of the server of 1000 Gbit. This means that when one second has passed, the first customer has a claim of 350. At the same time, a new customer arrives with a claim of 1350. The first customer however has higher seniority on her remaining claim of 350. This implies that although

the amount of the first customer's claim decreases, the strength on the claim increases compared to the new customer. Specifically, the strength of the first customer's claim is twice that of the customer that just arrived.

Since the service requirement of new customers exceeds the available resources presented by the server at every second, we can already predict that the queue will start growing infinitely (we have not imposed limits on customers patience). Thus, this queueing problem gives rise to an infinite series of fair division problems. For the first three, the Broomean problems can be written as:

$$\mathcal{B}_1^C = \left\langle 1000, \{1\}, 1350, 1 \right\rangle,$$
$$\mathcal{B}_2^C = \left\langle 1000, \{1,2\}, (350, 1350), \left(\frac{2}{3}, \frac{1}{3}\right) \right\rangle,$$
$$\mathcal{B}_3^C = \left\langle 1000, \{1,2,3\}, (8.5, 691.5, 1350), \left(\frac{3}{6}, \frac{2}{6}, \frac{1}{6}\right) \right\rangle,$$

or as represented by Table 6.3 which also includes the allocations by the absolute priority rule. As evident from Table 6.3c, the weighted proportional rule will once again assign a larger share of the resource than the first customer has a claim to (it assigns 9.3 when the claim is only 8.5). The absolute priority rule accommodates this. It is worth noting here that after the first second, the first customer is assigned a rather large share of its claim in comparison to the second customer. This is due to its earned seniority.

Table 6.3: Example C

(a) Initial allocation

i	Claims a_i	Strength s_i	Allocation by $P^\dagger$
1	1350	1	1000

(b) After one second

i	Claims a_i	Strength s_i	Allocation by $P^\dagger$
1	350	2/3	341.5
2	1350	1/3	658.5

(c) After two seconds

i	Claims a_i	Strength s_i	Allocation by P	Allocation by $P^\dagger$
1	8.5	3/6	9.3	8.5
2	691.5	2/6	501.3	501.7
3	1350	1/6	489.4	489.8

Notes: In this example, customers arrive with one second intervals and all have a claim with an amount of 1350 all with equal urgency. Although part of the claim is satisfied, the strength of the remaining claim increases due to its seniority. The queue will grow indefinitely since every new customer arrives with a service requirement that exceeds the available resources at every epoch. This table represents the first three epochs.

6.4 Taking stock

A few general learnings can be drawn from these examples. First and foremost, the examples show that queues are *dynamic*. They may either update, because the server does not have the resources to serve all customers immediately, as seen in Examples A and B. Or the dynamic component may be caused by the continuous arrival of customers over the course of time, as shown in Example C. In both cases, one is faced with a new problem of fair division at every epoch that can be represented as a new Broomean problem. The dynamic component provides another level of complexity to the analysis of fair division problems which are not present when dealing with e.g., bankruptcy problems. So far, I've implicitly assumed that it is morally relevant to consider fairness in every epoch but this assumption is not universally true by necessity. As Maccio, Hogg and Down (2018, p. 139) notes, securing fairness at every point in time does not necessarily lead to an overall fair system. Their conclusion, however, depends heavily on the definition and measurement of fairness, one adopts.

For all examples, I assumed that the server can serve multiple customers at the same time. That is, service is divisible between customers. This makes the proportional satisfaction of claims easy. In other queueing systems, however, the server can only serve one individual at a time. So service (at a given period in time) is indivisible. On Broome's account of fairness, whenever a good is indivisible it should be allocated by a lottery with weights equal to the strength of claims (Broome 1990, p. 89). But it may alternatively be suggested that the good should be given to the person with the largest claim. I leave questions regarding what to do when service is indivisible for future work.

There are infinitely many ways to assign the resources of a server amongst a number of claimants. The absolute priority rule of Wintein and Heilmann (2024a), which operationalises the notion of fairness that I adopt here, provides a function for determining the fair allocation of a server's resources. In order to assess the fairness of real applications, it is necessary to define a measure of deviation from this function. This is what I turn to now.

7 Measuring Fairness in Queueing

I've spent the previous sections describing what determines a fair allocation and how to conceptualise fairness for queueing. The last step is to define a *measure* of fairness of queueing that allows for measuring and comparing the fairness of different queueing disciplines. The measure I'm proposing is actually not so much a measure of fairness, as it is a measure of *un*fairness. I have argued that the absolute priority rule determines the fair allocation, and the fairness measure measures the deviation between the fair allocation and the actual allocation. That is, I'm measuring how unfair a given queueing discipline[23] is compared to the ideal situation as determined by the absolute priority rule.

Note, that by defining a measure, I'm allowing for direct and general comparisons of allocations. This is not a trivial matter. For reference, Wintein and

[23]I'm saying 'queueing discipline' but actually the measure is assessing the fairness of an actual, realised disposition of service to a specific set of customers. To avoid this cumbersome expression, I simply talk of the fairness of a queueing discipline, thereby disregarding deviations in realised disposition of customers. See also my mention of queueing disciplines as 'intended dispositions' in Section (2.2).

Heilmann (2024a, 8f.) apply only a minimal criteria for comparing allocations and claim satisfaction: they do comparisons only in a component-wise and ordinal manner. Assume that two individuals have claims of 10 each. The criteria of comparison of Wintein and Heilmann (2024a) allows them to conclude that the allocation $(5, 6)$ satisfies claims to a greater extent than the allocation $(5, 5)$. The measure of comparison is however inconclusive with regards to comparing allocations where one individual receives more while another receives less, such as $(8, 3)$ and $(7, 4)$.

This won't suffice for the purposes of a measure of fairness. However, providing an exact measure of fairness involves making choices on e.g., how to value few but large deviations from the fair allocation, compared to many but small deviations. It is important to realise that the choice of mathematical operators on measuring dispersion comes with philosophical commitments. How one measures dispersion implicitly means taking a stance with regards to distributional concerns. In the following, I make these philosophical commitments transparent, and my development of the measure thereby distinguishes itself from other measures in the literature. It is my hope that this transparency allows those who hold different views with regards to distributional concerns to make different choices.

The fairness measure, I propose, measures the deviation from the fair allocation. The deviation consists of two parts: 1) a measure of (total) difference from the fair allocation, or 'discrimination' as I shall dub it, and 2) a measure of dispersion of these differences. As it turns out, the two parts nicely maps the two dimensions — absolute and comparative — of fairness, I have described in Section 4. In this section, I first introduce and explain discrimination. I then discuss the appropriateness of different measures of dispersion and argue in favour of the mean absolute deviation operator. Finally, I combine the two elements to the Broomean Fairness Measure for Queueing.

7.1 Discrimination

In Section 4, I argued that fairness requires I) that claims are satisfied to the greatest possible extent as long as no one receives more than they have a claim to, II) that claims are satisfied in proportion to their strength and III) that priority is given to the first criteria whenever it conflicts with the second. The *absolute priority rule* $P^\dagger$ operationalises this notion of fairness and provides an allocation accordingly. So for a given queueing problem that can be represented by the Broomean problem $\mathcal{B}$, the rule demands that an amount $P^\dagger(\mathcal{B})_i$ of the resource is allocated to individual i. So, $P^\dagger(\mathcal{B})_i$ is individual i's 'fair share' of the resource.

Let the amount of the resource that is actually allotted to the individual be given by s_i. Then, I define the 'individual discrimination' as $d_i \equiv s_i - P^\dagger(\mathcal{B})_i$. Individual discrimination can be either positive, corresponding to a situation where the individual in questions receives more than her fair share, or negative, corresponding to a situation where the individual is allotted less than her fair share. Accordingly, I define 'total discrimination' as the sum of individual discrimination for all customers, i.e. $\mathcal{D} \equiv \sum_{\forall i} d_i$.

I use the term 'discrimination' for consistency with Raz, Levy and Avi-Itzhak (2004). It can be thought of as an 'offence' to fairness. The notion is identical to that of Raz, Levy and Avi-Itzhak (2004) in the sense that it denotes the

individual difference between the fair share of the resource and the share of the resource that is actually allotted. However, my notion of discrimination differs with regards to what defines the fair share. Whereas I argue that the fair share is determined by the absolute priority rule, Raz, Levy and Avi-Itzhak (2004) argue that fairness requires that every customer receives an *equal* share. Whenever I refer to the 'fair share' in the following, I'm referring to the share dictated by the absolute priority rule.

It is worth noting that the total discrimination has the property that whenever the server is using its full resource (in other words, the good is exhausted, or the server is 'non-idling'), every positive discrimination will be countered by a negative discrimination. To see this, consider the following example.

Example 7.1. There is a server with a total resource of 200 and two customers that have claims with equal amounts and equal strength. Both customers have claims taking the value 100. Fairness, as operationalised by the absolute priority rule, thus requires that they're assigned the same amount of the server's resources, i.e. $P^\dagger(\mathcal{B}^a)_1 = P^\dagger(\mathcal{B}^a)_2 = P^\dagger(\mathcal{B}^a) = 100$. Assume now that instead of being assigned their fair shares, the first customer is assigned $s_1^a = 50$ and the other is assigned $s_2^a = 150$ (since $s_1^a + s_2^a = 200$, the server is non-idling). Then the individual discrimination for each customers is given by

$$d_1^a = s_1^a - P^\dagger(\mathcal{B}^a) = 50 - 100 = -50$$
$$d_2^a = s_2^a - P^\dagger(\mathcal{B}^a) = 150 - 100 = 50,$$

which implies that the total discrimination is $\mathcal{D}^a = d_1^a + d_2^a = 0$.

Because the good is exhausted, total discrimination will be $D = 0$. Remember that on my use of the term, fairness has both an absolute and a comparative dimension. Total discrimination tells us something the *absolute* unfairness of the allocation. Whenever total discrimination is $\mathcal{D} = 0$, the good is exhausted and the absolute dimension of fairness has been addressed. In contrast, the following alteration of Example 7.1 describes a situation where the resource is not exhausted.

Example 7.2. Consider the same scenario as Example 7.1, but assume instead that the server only allocates half its resources (i.e., 100). The relative shares of the total resource allocated to each individual remain the same such that $s_1^b = 25$ and $s_2^b = 75$. By order of the good-exhaustion principle, the absolute priority rule still demands that the full resource is allocated, and thus recommends an allocation $P^\dagger(\mathcal{B}^b)_1 = P^\dagger(\mathcal{B}^b)_2 = P^\dagger(\mathcal{B}^b) = 100$. This implies that the individual discrimination factors become:

$$d_1^b = s_1^b - P^\dagger(\mathcal{B}^b) = 25 - 100 = -75$$
$$d_2^b = s_2^b - P^\dagger(\mathcal{B}^b) = 75 - 100 = -25,$$

which means that the total discrimination becomes negative, that is $\mathcal{D}^b = d_1^b + d_2^b = -100$.

So because the good is not allocated fully, total discrimination is negative. If we wanted to, we could choose to rename 'total discrimination' as '(indicator of)

absolute (un)fairness'. I'll abstain from doing so, to upkeep a terminological connection to *individual* discrimination, and to maintain that total discrimination is *one* indicator, or measure, of absolute unfairness (in theory, we could imagine other indicators). Total discrimination thus maps the absolute dimension of fairness and provides an indicator for assessing whether the good-exhaustion principle (cf. Section 4.4) is met. I should clarify that the mapping is not meant to be a one-to-one correspondence with the good-exhaustion principle from the fairness formula, as the good-exhaustion principle requires that no individual is assigned more than they have a claim to. By contrast, the notion of discrimination is only relevant because there are individuals who receive more than what fairness suggests, and this also includes potentially receiving more than they have a claim to.

Bear in mind, that just because total discrimination is $\mathcal{D} = 0$, it does not mean that all claims are satisfied. Far from it. Rather, it means that claims are satisfied to the greatest extent possible, *given* the available resource. In an extreme scenario, where there is no resource available, the absolute priority rule will demand that every individual is given nothing, because there's nothing to give. Giving nothing to everybody, will imply that total discrimination is zero. But clearly, no claims have been satisfied (assuming that some individuals have non-zero claims). It is thus important to be aware that discrimination is measured relative to the allocation determined by the fairness formula — not relative to claims, or some other distributive principle such as equality.

7.2 Measure of dispersion

The value of total discrimination depends on whether the good is exhausted or not, and thus informs us about the satisfaction of the absolute fairness of a distribution. However, it does not say anything about the comparative fairness of the allocation. As long as the resource is exhausted, total discrimination could be $\mathcal{D} = 0$, regardless of whether all of the resource is allocated to one individual with only little claim to the resource, or the resource is allocated such that every individual receive their fair share. In order to assess the comparative fairness of the allocation one must inspect the dispersion of the values of individual discrimination. To examine the properties of dispersion measures, I assume in this section that the good is exhausted, such that $\mathcal{D} = 0$.

Multiple measures of dispersion exist. Here, I treat three measures — range, variance and mean absolute deviation — each of which emphasises different aspects of the dispersion of the distribution and thus maps different distributive principles. I conclude that the mean absolute deviation is the most appropriate measure of dispersion for these purposes.

7.2.1 Range

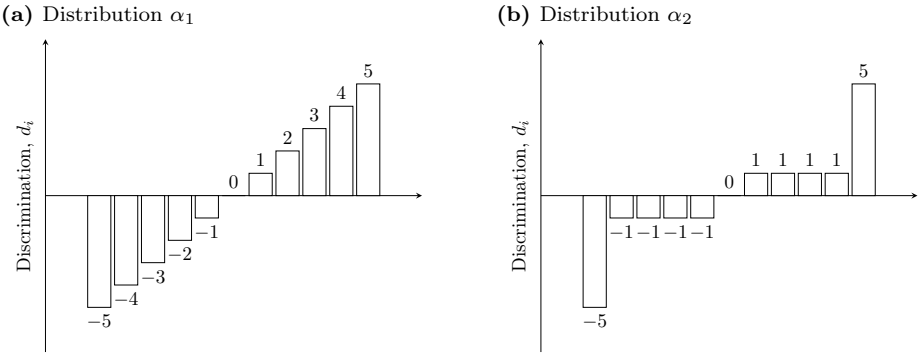
The range of a distribution is measured as the difference between the largest and the smallest value of the distribution. It thus only considers the extreme values of individual discrimination and disregards the rest. As such, the extreme values of the distribution is given absolute priority. Consider the two distributions illustrated in Figure 7.1. Both subfigures display distributions of individual discrimination for 11 customers ordered by magnitude of discrimination values, ranging from negative to positive. In both cases, there are five customers who

receive less than their fair share and five customers that receive more than their fair share. Due to this symmetry of the values of individual discrimination, the reader may confirm that total discrimination is $\mathcal{D} = 0$.

Notably, the range of both distributions is identical[24], namely 10. I believe, however, that most people would intuitively argue that distribution α_1 is more unfair compared to distribution α_2. The reason is that we can arrive at distribution α_2 by making a number of pure (i.e., no resource is wasted when making the transfer), gap-diminishing transfers of the resource from one customer to another, leaving others unaffected. To see this, number each customer from lowest to highest, such that the person with the lowest value of discrimination is customer 1, and the person with the highest value of discrimination is customer 11. Then take 3 units of the resource from customer 10, who has a positive value of discrimination of 4, and give them to customer 2. This constitutes an improvement from the point of fairness because customer 10 is receiving more than her fair share, whereas customer 2 is receiving less. Thus, taking 3 units of the resource from customer 10 does not harm her *from a point of fairness*. Obviously, there may be other reasons that taking 3 units of the resource from customer 10 is harmful to her. For instance, it might be against her preference. But as I'm solely concerned with fairness, this is not a worry I consider.

Moving on, one can take 2 units from customer 9 and give to customer 3, and further take 1 unit from customer 8 and give to customer 4. Doing so, it is possible to move from distribution α_1 to distribution α_2 without changing the order of the customers[25]. But despite these fairness-improving transfers, the range-operator supposes that both distributions are equally fair because the range of the distributions are equal. Therefore, I do not think that range is an appropriate dispersion measure of comparative fairness, and shall therefore disregard it going forward.

Figure 7.1: Distributions of discrimination, α_1 and α_2



[24]The range of a distribution, given by the vector ω, is given as as $R(\omega) \equiv \max\omega - \min\omega$. In this case, we can write the distributions on vector form as $\alpha_1 = (-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5)$ and $\alpha_2 = (-5, -1, -1, -1, -1, 0, 1, 1, 1, 1, 5)$. Then $R(\alpha_1) = 5 - (-5) = 10$ and $R(\alpha_2) = 5 - (-5) = 10$, hence $R(\alpha_1) = R(\alpha_2)$.

[25]In the social welfare literature, a variant of this argument with respect to well-being is known as the Pigou-Dalton axiom. See for instance Adler (2019, p. 274).

7.2.2 Variance

An alternative and commonly used measure of dispersion is *variance*. Raz, Levy and Avi-Itzhak (2004) use variance as measure of dispersion for the RAQFM. The variance is a measure of the spread of a distribution around its mean and computed as the expected value of the squared deviation from the mean of a variable. Let d denote the discrimination of an arbitrary customer, then the variance is given as

$$\mathrm{Var}(d) \equiv E[d - E[d]]^2.$$

Whenever total discrimination is $\mathcal{D} = 0$[26], the mean of individual discrimination is zero, i.e. $E[d] = 0$, and the variance simplifies to $\mathrm{Var}(d) = E[d^2]$. In contrast to the range, the variance takes all values in a distribution into account. Consider again the distributions depicted in Figure 7.1. The variance of distribution α_1 is $\mathrm{Var}(\alpha_1) = 10$, whereas for distribution α_2 it is $\mathrm{Var}(\alpha_2) = 5.27$, thus confirming the intuition that distribution α_2 is comparatively more fair than distribution α_1.

An important aspect of variance is that because it takes the expected value of the squared value, it assigns a larger weight to large values of discrimination. Consider the distributions depicted in Figure 7.2. In both cases, there are 5 customers, once again ordered by the value of their individual discrimination, ranging from negative to positive. In both distributions, there is one customer that receives her fair share, i.e. the share recommended by the absolute priority rule, so the value of her individual discrimination is 0. At the same time, there are two customers who are receiving more than their fair share, and two customers who receive less than their fair share of the resource. Importantly, the total amount of the resource given to customers in excess of their fair shares, i.e. the sum of positive individual discrimination, is the same in both distribution β_1 and distribution β_2, namely $5 + 1 = 3 + 3 = 6$. And by symmetry, the sum of all negative discrimination is the same for both distributions, namely $-5 - 1 = -3 - 3 = -6$.

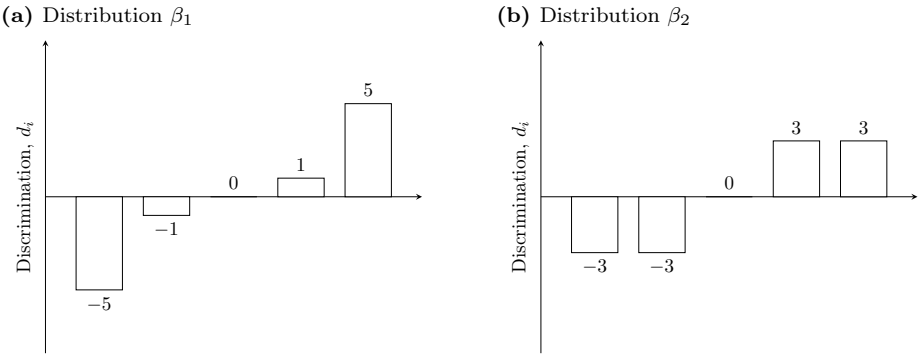
Because the variance puts emphasis on large deviations from the mean, the variance of distribution β_1 is larger than the variance of distribution β_2. Specifically, $\mathrm{Var}(\beta_1) = 10.4$ and $\mathrm{Var}(\beta_2) = 7.2$. If the variance is used as a measure of comparative fairness, the conclusion is that distribution β_2 is more fair than distribution β_1. In contrast to the range-operator that gave *absolute* priority to the extreme values and fully neglected the values in between, variance gives *relative* priority to larger values. The comparison of the distributions β_1 and β_2 show that large deviations from the mean tend to drive up the variance and thus determine distributions with comparatively few, but large deviations as overall more unfair than distributions with comparatively many, but smaller deviations.

7.2.3 Mean Absolute Deviation

The *mean absolute deviation* (MAD), or average absolute deviation, does not place such special weight on large values of discrimination. It measures the mean

[26]Remember, that $D = \sum_{i=1}^{n} d_i$, where n is the number of customers in the distribution. Then it can easily be confirmed by considering the population mean that $E[d] = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}D = 0$.

Figure 7.2: Distributions of discrimination, β_1 and β_2



of the absolute deviations from some central measure such as the arithmetic mean, $E[d]$. Let once again d be the discrimination of some arbitrary customer. Then, the mean average deviation is given by:

$$\text{mad}(d) \equiv E[|d - E[d]|],$$

where $|d - E[d]|$ denotes the absolute deviation from the mean. Whenever total discrimination is $\mathcal{D} = 0$, it follows as before that $E[d] = 0$, and thus that $\text{mad}(d) = E[|d|]$.

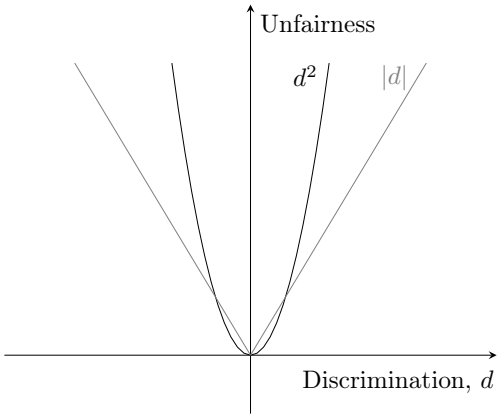
Consider again the two distributions β_1 and β_2 of Figure 7.2. Whereas the variance of the distributions were such that Var(β_1)>Var (β_2) because of the large deviations in distribution β_1, the MAD puts equal weight on all values of discrimination regardless of how large their values. This implies that the MAD for both distributions is the same, such that mad (β_1) = mad (β_2) = 2.4. This is a result of the fact that the sum of all positive and negative discriminations, respectively, are equal in both distributions.

7.2.4 Comparing variance and MAD

Comparing the variance and the mean absolute deviation operators, it shows that both operators work by taking the average of some function, or transformation, of individual discrimination (still assuming that $\mathcal{D} = 0$). In Figure 7.3, I have plotted these transformation functions, or 'elementary functions' as I shall refer to them. Here, discrimination is plotted along the x axis, rather than along the y axis as in Figures 7.1 and 7.2. For every value of discrimination d, a corresponding value of unfairness is plotted along the y axis[27]. The graph thus shows how different values of discrimination translates into levels of unfairness according to the two different measures. The overall (comparative) unfairness of a system is then obtained by taking different values of discrimination, reading off the individual values of unfairness that they translate into, and finally averaging the individual contributions. The convictions held with regards to how individual discrimination should be translated into values of unfairness determines whether one chooses one elementary function over the other.

[27] An equivalent representation would have fairness (and not *un*fairness) along the y axis in which case the graph would simply be inverted.

Figure 7.3: Variance, mean average deviation, and unfairness



Comparing d^2 (elementary function for variance) to $|d|$ (elementary function for MAD), it becomes evident that the increase in unfairness from a one unit increase in discrimination is constant when regarding the MAD (as long as the change is on the same side of the y axis, of course). For variance, however, the size of the increase in unfairness from a one unit increase in discrimination however depends on the level of discrimination one is at. That is, not only is the unfairness greater the farther away from the fair share you are, but the *rate* at which the level of unfairness changes increases the farther away from the fair share you are. Using variance as measure of unfairness thus implies evaluating large values of discriminations as morally more important than small values of discrimination. It is *more than* twice as bad to be twice as far from obtaining one's fair share. It implies that the largest improvements to the overall fairness of an allocation arises by transferring resources to or from individuals far from their fair share. When using MAD as a measure of unfairness, it only matters that transfers reduce the total sum of discriminations.

I believe many people would *prima facie* find it intuitively appealing that the dropoff in fairness from a unit change in discrimination increases the farther you are from your fair share[28]. However, I have two reasons for disregarding this intuition. First, I conjecture that many people hold the intuition based on some believed congruence with other philosophical principles of distribution. This connection is not necessarily existent, I argue. Second, I argue that measures, where the change in fairness from a unit change in discrimination depends on the level of discrimination, are subject to manipulation by groups.

To see that the intuition is not necessarily congruent with other philosophical principles of distribution, it is important to remember that an allocation with no individual discriminations is not necessarily a situation of equality of resources. Discriminations are differences between the actual allotted resource and the amount of resource as demanded by the absolute priority rule. How much of the

[28]Shelly Kagan (2015, p. 227) has done elaborate work with regards to desert, a moral notion closely related to fairness, that deserves mentioning here (pun intended). He argues that it is intuitively appealing that the dropoff in value produced by each additional unit change in well-being increases the farther you are from your peak. To my awareness, Kagan (2015) does not provide any foundation for this intuition (nor does he do so in earlier work, such as Kagan (2005)).

resource, the rule suggests to allocate to a particular individual, depends on the individual's claims. Imagine a situation in which two individuals are allotted the same amount of the resource, but the discrimination of one individual is much larger because the individual has a larger claim. So despite both individuals receiving the same amount of the resource, there is discrimination. So, if the intuition is based on concerns of equality of resources — as I believe concerns of equality most often are — then this intuition is flawed.

Alternatively, the intuition may be grounded in concerns of priority. The priority view suggests that it is morally more important to benefit the people who are worse off, not because it reduces inequality but because it benefits those worse off more (Parfit 2000). With regards to fairness, we may interpret the priority view as saying that it is morally more important to reduce the discrimination, or 'offence' to fairness, for those who are being treated unfairly. Despite its persuasiveness with regards to resources, I believe the underlying argument of the priority view fares less well in regards to discrimination and fairness. The view may explain why we should allocate resources to those worse off. It may also possibly explain why we should care about individuals who receive less than their fair share. But I do not believe the view explains convincingly why we should care equally about reducing positive discrimination as about reducing negative discrimination, as e.g. a variance measure implicitly assumes. So, I do not believe that concerns of priority give satisfying grounds for the intuition for assigning a relative larger importance to large values of discrimination. And in turn, I do not find compelling evidence that the intuition is congruent with other philosophical principles of distribution.

The second reason for disregarding the intuition has to do with manipulability, or lacking strategy-proofness. Whenever measures of comparative fairness rely on elementary functions where the change in fairness from a unit change in discrimination also depends on the current level of discrimination, the measure becomes susceptible to manipulation. It means that by grouping their claims, customers can make a scenario seem more or less unfair than it would be before the grouping. Table 7.1 reproduces the distributions β_1 and β_2 from Figure 7.3. In addition, I include a variant distribution β_3 in which four individuals (individuals I and II, as well as individuals III and IV) have paired up, such that their total discrimination is -6 and 6, respectively.

Evidentially, comparative fairness measured by variance is larger in distribution β_1 compared to distribution β_2, as discussed above. The comparative fairness as measured by the mean absolute deviation (MAD) is the same in both scenarios, since the measure is not sensitive to larger values of discrimination, opposed to variance. Similarly, the mean absolute deviation of distribution β_3 is also the same as distributions β_1 and β_2. But the comparative fairness — if measured by the variance — increases due the formation of groups[29], despite no changes have been incurred to individuals claims or their allocations as such.

Taken together, these are the reasons for advocating the mean absolute deviation over variance as a measure of comparative unfairness. The argument by Raz, Levy and Avi-Itzhak (2004) that the MAD is mathematically inconvenient

[29]Note, that average values are still computed based on the total number of *individuals* despite groups having formed. For instance the variance is computed as $\text{Var}(\beta_3) = \frac{1}{5}((-6)^2 + 0^2 + 6^2) = \frac{1}{5} \cdot 72 = 14.4$. Obviously, if we computed the average based on number of separate *entities* (in the example there are three entities, but five individuals), the mean absolute deviation would also be affected by grouping.

Table 7.1: Variance compared to mean absolute deviation

Individual	Distribution β_1	Distribution β_2	Distribution β_3
I	-3	-5	-6
II	-3	-1	
III	0	0	0
IV	3	1	6
V	3	5	
Mean	0	0	0
MAD	**2.4**	**2.4**	**2.4**
Variance	**7.2**	**10.4**	**14.4**

Notes: The table shows distributions β_1, β_2 and β_3 that both have the same sum of positive and negative discrimination. Distributions β_1 and β_2 are also depicted in Figure 7.3. It shows that the variance is larger for distribution β_2 than for distribution β_1 due to larger values of discrimination. In distribution β_3, two pairs of individuals have grouped their claims such that their joint discrimination is aggregated. Once, again the mean absolute deviation is unaffected whereas the variance increases.

is unfounded[30] as well as ethically irrelevant.

7.2.5 Implicit assumptions

Figure 7.3 and the arguments above encompasses a few implicit assumptions that ought to be made explicit. First, I treat instances of positive and negative discrimination symmetrically as I consider to be required from the proportionality requirement of fairness. The principle that claim amounts are satisfied in proportion to their strength is equally relevant, regardless of whether the actual allocation assigns resources above or below an individual's fair share. People who are less committed to the symmetric treatment of positive and negative discrimination could choose different 'elementary functions' as basis for their unfairness measures. For instance, a measure like $E[e^{-x}]$ could be considered more in line with the priority view described above. Such a measure will give large weight to negative values of discrimination and assign decreasing weights to positive values of discrimination.

Second, I'm implicitly assuming that overall fairness is affected in the same way regardless of which individual is being treated unfairly. This can be thought of as some sort of 'anonymity' assumption, that it does not matter who is treated unfairly. This includes the principle that individuals who have larger claims have the same elementary function as individuals who have larger claims. This is implicit in choosing discrimination, and not the overall level of resource to be allocated say, on the x axis[31].

Third, the discussion above should allow readers with convictions that differ from mine to apply other 'elementary functions' when assessing comparative fairness. However, I do believe that my arguments against variance hold for

[30] If not able to compute the absolute value directly, one may compute, assuming that $d \in \mathbb{R}$ among the real numbers, it as $|d| = \sqrt{d^2}$.

[31] In his work on desert, Kagan (2015) chooses this alternative approach in order to allow for option that it is morally more important that morally significant individuals get what they deserve in comparison to others.

any function that assigns more weight to large values of discrimination, i.e. any function $f(d) = d^{\xi}$ for $\xi > 1$. In principle, one could also imagine an elementary function that assigns lower weight to large values of discrimination, i.e., $\xi < 1$. I have not discussed these here, since I do not see any compelling reasons to consider such functions.

7.3 The Broomean Fairness Measure for Queueing

I have so far only considered the distance from the fair allocation in single epochs. However, as demonstrated in Section 6, queues are dynamic: Servers have a set amount of resources in each period, and customers can be assigned resources in multiple periods. When assessing the fairness of an allocation, we must therefore consider both the aggregated amount of resources designated by the fair rule and the amount of resources actually allotted over the period of time the customer is in the queue.

Assume that time is discrete, $t = 1, 2, 3, \ldots$. Then, the resources that are actually allotted to individual i are given by $s_i(t)$. As seen in Section 6, every epoch gives way to a new Broomean problem $\mathcal{B}(t)$[32], hence discrimination of individual i for epoch t is given as:

$$d_i(t) = s_i(t) - P^{\dagger}(\mathcal{B}(t))_i. \tag{7.1}$$

Given an arrival time b_i and a departure time d_i, the aggregated discrimination of individual i is defined as:

$$D_i \equiv \sum_{i|t\in(b_i,d_i)} d_i(t), \tag{7.2}$$

Accordingly, I redefine the total discrimination of all individuals aggregated over time as:

$$\mathcal{D} \equiv \sum_{\forall i} D_i. \tag{7.3}$$

A *scenario* is defined analogous to Avi-Itzhak, Levy and Raz (2007, p. 66) and describes a specific pattern of arrival, departure times and allotted shares of the server capacity for all individuals $i \in N$, denoted as $S = \{a_i, d_i; s_i\}_{i\in N}$. The unfairness ϕ^S of a given scenario S can then be measured as the mean absolute deviation of individual discriminations, i.e.:

$$\phi^S = \frac{1}{n}\sum_{i=1}^{n} |D_i|, \tag{7.4}$$

where n is the number of customers in the set of customers N, and where I assume that the principle of good-exhaustion (cf. Section 4.4) is met — that is to say, the server allocates its full capacity — such that $\mathcal{D} = 0$. I denote ϕ^S the *Broomean Fairness Measure for Queueing (BFMQ)*.

[32]Note that this notation is equivalent to the one used in Section 6 $\mathcal{B}_t \equiv \mathcal{B}(t)$.

7.3.1 An improved measure

The BFMQ provides improvements to existing fairness measures for queueing, such as the RAQFM, in four ways as promised in Section 2. First, it generalises the assertion that all customers should receive equal shares of the resource. The assertion can simply be considered a simplifying assumption but it is important to recognise it as such. Second, it takes account of these differing service requirements in its assignment of resources to individuals. Third, it applies a more appropriate measure of dispersion and lays out the philosophical commitments related to it. And fourth, it is not based on arbitrary 'principles of fairness', but is founded an a rigorously defined account of fairness. So contrary to the belief of Avi-Itzhak, Levy and Raz (2008, p. 503), I believe that fairness measures do benefit from being grounded in 'formal conception[s] offered by 'deep thinkers".

8 Conclusion

In this thesis, I argue that there are occasions where fairness is an important aspect to consider when choosing between queueing disciplines. Failing to consider fairness due to a one-sided focus on queueing effectiveness may lead to unfair treatment of certain customers depending on the chosen discipline.

I suggest that certain queueing problems can be regarded as a problems of fair division, for instance when one regards computing power as a resource to be allocated amongst a number of customers in a queue, each of whom has a claim. I adopt the account of fairness presented by Wintein and Heilmann (2024a; 2024b) and propose that queueing problems can be represented as Broomean problems. I argue that a customer's service requirement constitutes a claim for service with an amount equal to the resources required to satisfy the requirement, and that the strength of the claim depends on the customer's seniority and how urgent claim satisfaction is.

Given this representation, one can proceed to apply a fairness function in order to determine the fair allocation of the server's capacity. In Section 6, I provide examples of how a particular fairness function, *the absolute priority rule*, can be applied to a number of queueing problems and demonstrate how queueing adds another layer of complexity to the analysis of fair division problems by being *dynamic.*

In Section 7, I examine how the fairness of queues can be measured by introducing the notion of discrimination as the difference between the fair allocation and the actual amount of a resource allotted to a customer. I discuss different dispersion measures for assessing the comparative fairness of a distribution and argue that the *mean absolute deviation* is the most appropriate measure. The measure I propose, The Broomean Fairness Measure for Queueing (BFMQ), can be considered a generalisation of the Resource Allocation Queueing Fairness Measure (RAQFM) of Raz, Levy and Avi-Itzhak (2004) and I argue that the BFMQ improves the measure along a number of dimensions.

I do not provide a detailed examination of the bounds and properties of the BFMQ, but due to its resemblance with the RAQFM, many of the properties and bounds will be similar. Future work should prove and examine these properties. Until then, I refer to the work done by Avi-Itzhak, Levy and Raz (2007) on the RAQFM.

The emphasis of this project has been on providing the philosophical foundation for fairness measures that has been lacking in the literature on fair queueing. In doing so, I have chosen to delimit the project by only considering cases where the server's capacity is divisible and can be shared among a number of customers at the same point in time. It means that the server can serve multiple individuals at a time, at arbitrarily small increments in the amounts of the serving capacity. An obvious next step is to consider situations where only one customer can be served at a time, meaning the server's capacity is indivisible. Fair division of indivisible resources pose philosophical difficulties of their own. Sorting out these issues is a task I leave for future work.

References

Adler, Matthew D. (2019). *Measuring Social Welfare: An Introduction*. New York: Oxford University Press. 317 pp. ISBN: 978-0-19-064302-7 978-0-19-064303-4.

Aristotle (2009). *The Nicomachean Ethics*. Ed. by Lesley Brown. Trans. by William David Ross. ed. rev. Oxford World's Classics. Oxford: Oxford university press. ISBN: 978-0-19-921361-0.

Avi-Itzhak, Benjamin, Eli Brosh and Hanoch Levy (2007). "SQF: A Slowdown Queueing Fairness Measure". In: *Performance Evaluation* 64.9-12, pp. 1121–1136. ISSN: 01665316. DOI: 10.1016/j.peva.2007.06.018. URL: https://linkinghub.elsevier.com/retrieve/pii/S0166531607000685.

Avi-Itzhak, Benjamin and Hanoch Levy (2004). "On Measuring Fairness in Queues". In: *Advances in Appliled Probability* 36.3, pp. 919–936.

Avi-Itzhak, Benjamin, Hanoch Levy and David Raz (2007). "A Resource Allocation Queueing Fairness Measure: Properties and Bounds". In: *Queueing Systems* 56.2, pp. 65–71. ISSN: 0257-0130, 1572-9443. DOI: 10.1007/s11134-007-9025-x. URL: http://link.springer.com/10.1007/s11134-007-9025-x.

— (2008). "Quantifying Fairness in Queuing Systems: Principles, Approaches, and Applicability". In: *Probability in the Engineering and Informational Sciences* 22.4, pp. 495–517. ISSN: 0269-9648, 1469-8951. DOI: 10.1017/S0269964808000302. URL: https://www.cambridge.org/core/product/identifier/S0269964808000302/type/journal_article.

— (2011). "Principles of Fairness Quantification in Queuing Systems". In: *Network Performance Engineering: A Handbook on Convergent Multi-Service Networks and next Generation Internet*. Ed. by Demetres D. Kouvatsos. Lecture Notes in Computer Science 5233. Berlin Heidelberg New York: Springer, pp. 284–300. ISBN: 978-3-642-02742-0.

Broome, John (1984). "Selecting People Randomly". In: *Ethics* 95.1, pp. 38–55. JSTOR: 2380545. URL: http://www.jstor.org/stable/2380545.

— (1990). "Fairness". In: *Proceedings of the Aristotelian Society* 91, pp. 87–101. ISSN: 00667374, 14679264. URL: http://www.jstor.org.eur.idm.oclc.org/stable/4545128.

— (1991). *Weighing Goods: Equality, Uncertainty and Time*. 1st ed. Wiley. ISBN: 978-0-631-17199-7 978-1-119-45126-6. DOI: 10.1002/9781119451266. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119451266.

Casas-Méndez, Balbina, Vito Fragnelli and Ignacio García-Jurado (2011). "Weighted Bankruptcy Rules and the Museum Pass Problem". In: *European Journal of Operational Research* 215.1, pp. 161–168. ISSN: 03772217. DOI: 10.1016/j.ejor.2011.05.033. URL: https://linkinghub.elsevier.com/retrieve/pii/S0377221711004589.

Chun, Youngsub (2016). *Fair Queueing*. Studies in Choice and Welfare. Cham: Springer International Publishing. ISBN: 978-3-319-33770-8 978-3-319-33771-5. DOI: 10.1007/978-3-319-33771-5. URL: http://link.springer.com/10.1007/978-3-319-33771-5.

Cooper, Robert B. (1981). *Introduction to Queueing Theory*. 2d ed. New York: North Holland. 347 pp. ISBN: 978-0-444-00379-9.

Demers, Alan, Srinivasan Keshav and Scott Shenker (1989). "Analysis and Simulation of a Fair Queueing Algorithm". In: *SIGCOMM '89: Symposium proceedings on Communications architectures & protocols*, pp. 1–12.

Gorden, Ethel Sherry (1987). *New Problems in Queues: Social Injustice and Server Producton Mangement*. Ph.D. Dissertation. Massachusetts Institute of Technology, p. 436.

Holm, Sune (2023). "The Fairness in Algorithmic Fairness". In: *Res Publica* 29.2, pp. 265–281. ISSN: 1356-4765, 1572-8692. DOI: 10.1007/s11158-022-09546-3. URL: https://link.springer.com/10.1007/s11158-022-09546-3.

Hooker, Brad (2005). "Fairness". In: *Ethical Theory and Moral Practice* 8.4, pp. 329–352. ISSN: 1386-2820, 1572-8447. DOI: 10.1007/s10677-005-8836-2. URL: http://link.springer.com/10.1007/s10677-005-8836-2.

Kagan, Shelly (10th Mar. 2005). "The Geometry of Desert". The Lindley Lecture (The University of Kansas).

— (2015). *The Geometry of Desert*. Oxford University Press. 676 pp. ISBN: 978-0-19-023372-3.

— (2019). *How to Count Animals, More or Less*. First edition. Uehiro Series in Practical Ethics. Oxford: Oxford University Press. 309 pp. ISBN: 978-0-19-882967-6.

Kayı, Çağatay and Eve Ramaekers (2010). "Characterizations of Pareto-efficient, Fair, and Strategy-Proof Allocation Rules in Queueing Problems". In: *Games and Economic Behavior* 68.1, pp. 220–232. ISSN: 08998256. DOI: 10.1016/j.geb.2009.07.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0899825609001547.

Kirkpatrick, James R. and Nick Eastwood (2015). "Broome's Theory of Fairness and the Problem of Quantifying the Strengths of Claims". In: *Utilitas* 27.1, pp. 82–91. ISSN: 0953-8208, 1741-6183. DOI: 10.1017/S0953820814000259. URL: https://www.cambridge.org/core/product/identifier/S0953820814000259/type/journal_article.

Maccio, Vincent J., Jenell Hogg and Douglas G. Down (2018). "On Slowdown Variance as a Measure of Fairness". In: *Operations Research Perspectives* 5, pp. 133–144. ISSN: 22147160. DOI: 10.1016/j.orp.2018.05.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2214716018300071.

Nagle, J. (1987). "On Packet Switches with Infinite Storage". In: *IEEE Transactions on Communications* 35.4, pp. 435–438. ISSN: 0090-6778. DOI: 10.1109/TCOM.1987.1096782. URL: http://ieeexplore.ieee.org/document/1096782/.

Parfit, Derek (2000). "Equality or Priority". In: *The Ideal of Equality*. Ed. by Matthew Clayton and Andrew Williams. Houndmills, Basingstoke, Hampshire : New York: Macmillan Press ; St. Martin's Press, pp. 81–125. ISBN: 978-0-312-23017-3.

Piller, Christian (2017). "Treating Broome Fairly". In: *Utilitas* 29.2, pp. 214–238. ISSN: 0953-8208, 1741-6183. DOI: 10.1017/S0953820816000303. URL: https://www.cambridge.org/core/product/identifier/S0953820816000303/type/journal_article.

Rawls, John (1958). "Justice as Fairness". In: *The Philosophical Review* 67.2, pp. 164–194.

Raz, David, Benjamin Avi-Itzhak and Hanoch Levy (2006). "Fairness Considerations of Scheduling in Multi-Server and Multi-Queue Systems". In: *Proceedings of the 1st International Conference on Performance Evaluation Meth-*

odolgies and Tools - Valuetools '06. The 1st International Conference. Pisa, Italy: ACM Press, p. 39. ISBN: 978-1-59593-504-5. DOI: 10.1145/1190095.1190145. URL: http://portal.acm.org/citation.cfm?doid=1190095.1190145.

Raz, David, Hanoch Levy and Benjamin Avi-Itzhak (2004). "A Resource-Allocation Queueing Fairness Measure". In: *ACM SIGMETRICS Performance Evaluation Review* 32.1, pp. 130–141. DOI: 10.1145/1012888.1005704.

Sandmann, W. (2005). "A Discrimination Frequency Based Queueing Fairness Measure with Regard to Job Seniority and Service Requirement". In: *Next Generation Internet Networks, 2005*. Next Generation Internet Networks, 2005. Rome, Italy: IEEE, pp. 106–113. ISBN: 978-0-7803-8900-7. DOI: 10.1109/NGI.2005.1431654. URL: http://ieeexplore.ieee.org/document/1431654/.

Scanlon, T.M. (1975). "Preference and Urgency". In: *The Journal of Philosophy* 72.19, pp. 655–669. JSTOR: 2024630. URL: https://www.jstor.org/stable/2024630.

Shortle, John F. et al. (2017). *Fundamentals of Queueing Theory*. Fifth edition. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons. 1 p. ISBN: 978-1-118-94356-4 978-1-118-94353-3.

Tang, Shanjiang, Ce Yu and Yusen Li (2022). "Fairness-Efficiency Scheduling for Cloud Computing With Soft Fairness Guarantees". In: *IEEE Transactions on Cloud Computing* 10.3, pp. 1806–1818. ISSN: 2168-7161, 2372-0018. DOI: 10.1109/TCC.2020.3021084. URL: https://ieeexplore.ieee.org/document/9185017/.

Thomson, William (2003). "Axiomatic and Game-Theoretic Analysis of Bankruptcy and Taxation Problems: A Survey". In: *Mathematical Social Sciences* 45.3, pp. 249–297. ISSN: 01654896. DOI: 10.1016/S0165-4896(02)00070-7. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165489602000707.

Wierman, Adam (2011). "Fairness and Scheduling in Single Server Queues". In: *Surveys in Operations Research and Management Science* 16.1, pp. 39–48. ISSN: 18767354. DOI: 10.1016/j.sorms.2010.07.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S1876735410000048.

Wierman, Adam and Mor Harchol-Balter (2003). "Classifying Scheduling Policies with Respect to Unfairness in an M/GI/1 £". In: *ACM SIGMETRICS Performance Evaluation Review* Volume 31.1, pp. 238–249. DOI: 10.1145/885651.781057.

Wintein, Stefan and Conrad Heilmann (2020). "Theories of Fairness and Aggregation". In: *Erkenntnis* 85.3, pp. 715–738. ISSN: 0165-0106, 1572-8420. DOI: 10.1007/s10670-018-0045-1. URL: http://link.springer.com/10.1007/s10670-018-0045-1.

— (2021). "Fairness and Fair Division". In: Heilmann, Conrad and Julian Reiss. *The Routledge Handbook of Philosophy of Economics*. 1st ed. New York: Routledge, pp. 255–267. ISBN: 978-1-315-73979-3. DOI: 10.4324/9781315739793-23. URL: https://www.taylorfrancis.com/books/9781315739793/chapters/10.4324/9781315739793-23.

— (2024a). "How to Be Absolutely Fair Part I: The Fairness Formula". In: *Economics and Philosophy*, pp. 1–24. ISSN: 0266-2671, 1474-0028. DOI: 10.1017/S0266267123000408. URL: https://www.cambridge.org/core/product/identifier/S0266267123000408/type/journal_article.

Wintein, Stefan and Conrad Heilmann (2024b). "How to Be Absolutely Fair Part II: Philosophy Meets Economics". In: *Economics and Philosophy*, pp. 1–23. ISSN: 0266-2671, 1474-0028. DOI: 10.1017/S026626712300041X. URL: https://www.cambridge.org/core/product/identifier/S026626712300041X/type/journal_article.

A Appendix

A.1 Other conceptualisation of fairness in queueing

When considering the fairness of queues, it is not only important to have a clear definition of fairness, but also of how fairness should be conceptualised for queueing. In this thesis, I adopt the view of i.a., Raz, Levy and Avi-Itzhak (2004) that queueing can be interpreted as the allocation of a scarce resource — the capacity of the server — amongst a number of claimants. This enables the application of fair division theories from the fairness literature. However, other interpretations of queueing exist in the literature.

Slowdown Scholars, such as Avi-Itzhak, Brosh and Levy (2007), Maccio, Hogg and Down (2018) and Wierman and Harchol-Balter (2003), argue that fairness may be related to the *slowdown* of customers service requirements. Slowdown is the ratio of response time (waiting time plus service time) to service requirement. If the response time of job x is $R(x)$, and the service requirement of job x is $S(x)$, then slowdown is $R(x) = \frac{T(x)}{S(x)}$. For the full system, one may regard either the expected slowdown, or the variance in slowdown of different customers.

On the 'slowdown'-view, it is argued that a queueing discipline is fair if the slowdown is constant among jobs. It is fair if response time is proportional to service requirement. In other words, it is fair if long jobs wait for longer. The justification for slowdown measures is often based on reference to the service requirement principle of fairness (cf. Section 2.2.1) which for this reason is sometimes also denoted 'proportional fairness' (Maccio, Hogg and Down 2018, p. 133). As I argue in 3.1, this principle does not reflect correct aspects of fairness. For this reason, I also do not agree with slowdown as an appropriate conceptualisation of fairness for queueing.

Violations to seniority Gorden (1987) and Sandmann (2005) both conceptualise fairness in queueing on the basis that it is fair if jobs are satisfied according to their seniority. Gorden (1987) thus argues that the first-come-first-served (FCFS) discipline, where customers are served according to their arrival time, is the fairest discipline. She then measures (un)fairness as the number of violations to the order determined by FCFS, i.e., the number of times an individual is served earlier than what their seniority warrants. Sandmann (2005) also counts the number of times the queue is 'jumped' but improves on Gorden's measure by also counting the 'size' of the violation, i.e., how much time the queue jumper saves by violating seniority.

Although I agree that seniority is an important determinant of a customer's claim to service, I argue in Section 5 that there are other determinants that are important to take account for when determining the fairness of a queueing discipline. For this reason, I believe the conceptualisation of fairness for queueing, that relies on violations to seniority, as incomplete.

Social choice approach The social choice approach to queueing, considered in particular by Chun (2016), argues that an agent's order in a queue can be translated to utility via a transformation procedure. An agent gets higher utility

their rank in the queue. At the same time waiting incurs costs to their utility. The queueing problem then consists in determining the right order and the appropriate monetary transfers.

As I argue in Section 3.2, this type of queueing system regarded by Chun (2016) deviates from the type of queueing systems that I regard, in particular by allowing for monetary transfers. Allowing for monetary transfers comes with the benefit that it enables the possibility of optimising for both fairness and effectiveness. Indeed, the costs imposed on agents from e.g. optimising for effectiveness can be offset by a monetary transfer. But ultimately I find it quite a rare trait in queueing systems, and I therefore disregard it here.

A.2 The fairness of a queueing system

For queueing systems that enter steady state, it is possible to compute the unfairness of the *queueing system* for a given queueing discipline. The *state* of a queueing system is determined by the number of customers in the system. The system is in *steady state,* or statistical equilibrium, whenever the rate of transition from one state to another is equal to the rate of transition out of that state (Shortle et al. 2017, p. 74). In practice, it implies amongst other things that the number of individuals in the queue is not constantly growing. Rather, the rate of customers arriving equals the rate of customers leaving the queue for every state of the system.

In case the system is in steady state, let D be a random variable describing the discrimination of an arbitrary customer in the the system. That is, D is the stationary of discrimination D_i. Then, the BFMQ of a queueing discipline (or policy) P is given as:

$$\Phi^P = \mathrm{E}[|D|]. \tag{A.1}$$

Analogously the fairness, as measured by the RAQFM, of a queueing discipline for a queueing system in steady state is given by :

$$\Psi^P = E[D^2]. \tag{A.2}$$

A.3 Continuous time

The BFMQ can be adapted to continuous time by changing Equation (7.2) such that it becomes

$$D_i \equiv \int_{b_i}^{d_i} d_i(t)dt, \tag{7.2a}$$

where $d_i(t)$ is defined in Equation 7.1 and b_i and d_i are arrival and departure epochs for customer i. The other definitions remain the same.

Analogously for the RAQFM:

$$D_i^{\mathrm{RAQFM}} \equiv \int_{b_i}^{d_i} d_i^{\mathrm{RAQFM}}(t)dt. \tag{A.3}$$