London School of Economics and Political Science

Philosophical and Ethical Aspects of Economic Design

Philippe van Basshuysen

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, May 07, 2019

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I confirm that parts of the Introduction and Conclusion are based on a book review in *Review of Political Economy* (van Basshuysen (forthcoming)). Chapter 4 is based on a publication in *Games* (van Basshuysen (2017)).

I declare that this thesis consists of $41,757$ words.

Philippe van Basshuysen

Abstract

This thesis studies some philosophical and ethical issues that economic design raises. Chapter 1 gives an overview of economic design and argues that a cross-fertilisation between philosophy and economic design is possible and insightful for both sides. Chapter 2 examines the implications of mechanism design for theories of rationality. I show that non-classical theories, such as constrained maximization and team reasoning, are at odds with the constraint of incentive compatibility. This poses a problem for non-classical theories, which proponents of these theories have not addressed to date. Chapter 3 proposes a general epistemology of economic engineering, which is motivated by a novel case study, viz. the reform of a matching market for medical practitioners. My account makes use of causal graphs to explain how models allow encoding counterfactual information about how market outcomes change if the design of the market changes. The second part of the thesis examines ethical issues. In Chapter 4, I apply tools from matching theory to gain insights into the distribution of refugees, such as among countries of the European Union. There is an ethical trade-off between the fairness of matchings and their efficiency, and I argue that in this context, fairness is the morally weightier criterion. Chapter 5 deals with the ethics of kidney exchange. Against critics, I give two arguments for the implementation of kidney exchange programmes. The first argument is that they are instrumental in meeting a moral obligation, namely to donate effectively. The second is that kidney exchange may increase the motivation for altruistic donations, because the donation of one kidney may trigger > 1 life savings. The final chapter identifies questions for future research and it closes with some thoughts about the future trajectory of economic design.

Acknowledgements

I would like to thank my supervisors, Luc Bovens and Bryan Roberts, for their support and encouragement to pursue my pitiful ideas.

I thank my fellow PhD students, some of whom have become close friends: Tom Rowe, Chris Marshall, Todd Karhu, David Kinney, Christina Easton, Bastian Steuwer, James Nguyen, Mantas Radzvilas, Ko-Hung Kuan, Silvia Milano, Aron Vallinder, Nicolas Wuethrich, Chloe de Canson, Joe Roussos, Paul Daniell, Nicolas Cote, Goreti Faria, Margherita Harris, Adam White, James Wills, Charles Beasley, Sophie Kikkert, Fabian Beigang, Nicholas Makins, Dmitry Ananyev, and the rest of the bunch. Special thanks to those who wasted their time giving feedback on some of the chapters.

This thesis has benefited from discussions, for which I'm grateful, with Jason Alexander, Anna Alexandrova, Richard Bradley, Eric Brandstedt, Fadi Esber, Roman Frigg, Francesco Guala, Jurgis Karpus, Aki Lehtinen, Matthias Lücke, Uskali Mäki, Andrés Perea, Sandro Provenzano, Katie Steele, Alex Teytelboym, Alex Voorhoeve, Kate Vredenburgh and Philipp Wichardt. Thanks also for comments to anonymous referees of the journals *Games* and *Philosophy of Science*. The numerous remaining mistakes are my sole responsibility.

Thanks to my examiners, Erik Angner and Kai Spiekermann, who let me pass for some reason, and in particular for an interesting discussion.

For the undeserved love and support I receive every day I thank Kristel, as well as the rest of my German and Mexican families. Thanks to Sofia for being so lovely.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Some institutions evolve naturally over time; others are designed with specific goals in mind. As we shall see, well-designed institutions can save people's lives, and improve the lot of many more. Conversely, a badly designed institution can lead to catastrophic social consequences, and it can produce adverse incentives that may in turn threaten the institution.

The design of institutions, in particular of markets, is the subject of much recent work in economic theory and practice. I use "economic design" as an umbrella for these endeavours. Various recent Nobel Memorial Prizes have been awarded for contributions to economic design, making it a highly topical field in economics.[1] At the same time, even though it raises many philosophical and ethical questions, it has been relatively neglected by philosophers.

[1] In 2007, the award went to Leonid Hurwicz, Eric Maskin und Roger Myerson, "for having laid the foundations of mechanism design theory" (Information for the Public 2007, `https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2007/popular-economicsciences2007.pdf`), and in 2012 to Alvin Roth and Lloyd Shapley, "for the theory of stable allocations and the practice of market design" (Press Release 2012, `https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2012/press_02.pdf`). The 2014 Prize to Jean Tirole could also be seen as an award for economic design, broadly conceived, in the field of industrial organization. Going further back, in 1994 John F. Nash, Reinhard Selten and John Harsanyi were awarded for their contributions to game theory, which plays a vital role in economic design. The 1996 Nobel Prize went to James Mirrlees and William Vickrey, the latter of whom pioneered auction design. Furthermore, in 2005, game theorists Thomas Schelling and Robert Aumann received the prize. I bet that the broad field will continue to recruit laureates.

Moreover, moral philosophy and economic design have at times appeared as adversaries. This is especially the case in debates about the moral limits of markets, in which many economists hold liberal views about their reach, while some moral philosophers argue for a narrower domain.

This thesis makes a case for the view that a cross-fertilization between philosophy and economic design is possible and insightful for both sides. I seek to show that, first, economic design has the potential to produce novel insights into a variety of philosophical questions. These include the nature of institutions and of rationality, how knowledge about the social world can be gained, and how society should be organised. Second, ethics should form an integral part of economic design because many design decisions are subject to substantive ethical questions.

Here, I shall give an overview of economic design. I will sketch its history, and I will identify important developments within the field. My overview starts with the theory of mechanism design in the next section, whose origins lie in the controversy over the relative merits of centrally planned versus market economies. I shall show that this large-scale, theoretical focus on economic systems that was prevalent in the early days of mechanism design theory, successively got replaced by a more narrow focus on particular marketplaces. In section 1.2, I show how economic designers increasingly act as "engineers", or even "plumbers", who repair deficient markets or create new ones where they can be expected to produce more desirable outcomes. In discussing these developments, I also mention the philosophical questions that economic design raises. In section 1.3, I discuss the relationship between economic design and ethics. In particular, I argue that ethics should form an integral part of economic design. Section 1.4 gives a short summary, and section 1.5 provides an overview of the chapters to follow.

1.1 Mechanism Design

Leonid Hurwicz is the founding father of mechanism design, which provides the theoretical foundations of economic design. Hurwicz started his 1973 Richard Ely Lecture, "The Design of Mechanisms for Resource Allocation", describing how this field differs from more traditional economic analyses:

> "Traditionally, economic analysis treats the economic system as one of the givens. The term "design" in the title is meant to stress that the structure of the economic system is to be regarded as an unknown. An unknown in what problem? Typically, that of finding a system that would be, in a sense to be specified, superior to the existing one. The idea of searching for a better system is at least as ancient as Plato's *Republic*, but it is only recently that tools have become available for a systematic, analytical approach to such search procedures. This new approach refuses to accept the institutional status quo of a particular time and place as the only legitimate object of interest and yet recognizes constraints that disqualify naive utopias." Hurwicz (1973), p. 1

Hurwicz was interested in answering large-scale questions about economic systems, in particular, to decide the controversy over the relative merits of centrally planned versus market economies. By the time of his writing, this planning controversy had been raging for decades: Oskar Lange (1942) and Abba Lerner (1944) argued that a centrally planned economy could in principle replicate the efficient allocation of resources in a free market and could improve on the workings of the free market by correcting market failures. Other theorists, notably Friedrich Hayek (1935) and Ludwig von Mises (1935), argued on the contrary. Hayek famously asserted that central planning could lead to efficient resource allocations only if the planner possesses at least as much information

about the desires and resources of other agents as the market mechanism generates spontaneously, but that it is not in the interests of agents to reveal their private information (Hayek (1945)). However, the economic models available at the time of the debate accounted for economic systems only as mechanisms for the allocation of scarce resources, but not as mechanisms for communicating private information that is widely dispersed throughout the economy. Therefore, they ignored the incentives for conveying information that different mechanisms provide to agents (cf. Myerson (2008)). Because a precise mathematical treatment of incentives was lacking, the planning controversy remained largely inconclusive.

Hurwicz (1972) provided a theory that introduced incentive constraints, in addition to resource constraints. Institutions are modelled as mechanisms that determine how social decisions and allocations of goods depend on the actions of the individuals interacting through these institutions. Importantly, their actions can include conveying private information, e.g. about their endowments, or their preferences. It is assumed that the individuals are rational and game theory is used to predict the institutional outcomes. In this framework, individuals may be strategic, while adhering to the rules of the game.[2] For instance, they may lie about their preferences if this is in their best interest. This leads to the concept of *incentive compatibility*: a mechanism is incentive-compatible if it implements some predefined social goal in equilibrium, that is, it gives everyone incentives to act according to the social plan.[3]

This analysis of incentives laid the foundation of mechanism design theory, which subsequently evolved rapidly. As Makowski and Ostroy (1992) wrote, "around 1970, the issue of incentives surfaced forcefully, as if a pair of blinders

[2]Note that, by allowing strategic behaviour, the model delimits the knowledge and power of the social planner (e.g. Trockel (2002), p. 30). The model thus excludes dictatorial systems, in which the social planner could simply tell everyone what to do.

[3]The idea of incentive compatibility appears already in Hurwicz (1953), where he suggests that mechanisms with this feature be called "cooperative", or "self-interest cooperative" (p. 7). Hurwicz (1960) seems to have used "incentive-compatible" for the first time. The canonical paper is Hurwicz (1972).

were removed" (p. 14). Among the most notable subsequent results is, first, the *Revelation Principle*, which characterises the set of feasible social outcomes as those that are implementable by incentive-compatible mechanisms;[4] and second, *Maskin Monotonicity*, which is a property of implementable social choice rules and which together with a second property (called "no veto power") can be used to construct mechanisms that implement the choice rule in question (Maskin (1999)). I will make use of the concept of incentive compatibility in the next chapter, where I show that certain theories of rationality, which are popular especially among philosophers, conflict with incentive-compatible institutional design. I will argue that this presents a challenge for those theories. We will also be led to a fundamental question in social ontology, namely, what are institutions? I will try to show that mechanism design yields important insights into this question.[5]

Let's consider again the planning controversy. Mechanism design provides flexible tools for comparing the efficiency and incentives that different economic systems provide. For instance, Eric Maskin (2015) shows that Hayek's claims – that the market is the only informationally efficient and incentive-compatible mechanism – are true when there are a large number of buyers and sellers and no externalities.[6] However, if these assumptions are not met, there are mechanisms that generally improve upon the market.

Furthermore, Roger Myerson (2009) models how different kinds of incentive

[4]Several theorists discovered the revelation principle independently around the same time in the late 1970s, including Dasgupta et al. (1979) and Myerson (1979), among others.

[5]Recent work by Francesco Guala and Frank Hindriks has refocused attention on this question (see Hindriks and Guala (2015a) and Guala (2016)). I will suggest some amendments to their theory.

[6]The relation between Hayek's work and mechanism design is peculiar. On the one hand, mechanism design can be seen as a consequence of Hayek's treatment of knowledge and incentives, and Hayek seems to have anticipated crucial results from mechanism design. On the other hand, treating economic systems as variables in the problem of finding a superior system to the existing one must have been utterly unacceptable to Hayek, because it must mean for him to mess with the desirable, "spontaneous order" of the market economy. (Note, however, that Hayek scholars have debated whether he believed that spontaneous orders are superior to artificial orders. Caldwell (2000) rejects this interpretation, while Angner (2004, 2007) defends it.)

problems figure in unbridled capitalism and in centrally planned economies. Remember that mechanisms determine how allocations of goods depend on individual actions. These actions may include means of communication, but also other actions, such as exerting efforts that the social planner cannot reliably observe. The social planner wants to design incentives for the agents to report their private information honestly, otherwise she will face *adverse selection* problems. She also wants to design incentives for agents to do what they are supposed to do when their actions are hidden, otherwise there are *moral hazard* problems. Accordingly, we can distinguish two types of incentive constraints that social planners should take into account: informational incentive constraints to avoid adverse selection, and strategic incentive constraints to eliminate moral hazard.

The distinction between informational and strategic incentive constraints is significant for the planning debate: in Myerson's models, socialism can be shown to have advantages concerning adverse selection problems, but disadvantages concerning moral hazard. Capitalism, in contrast, shows disadvantages concerning adverse selection but advantages concerning moral hazard. It is not difficult to make intuitive sense of this result. For instance, motivating workers in the absence of significant monetary incentives and property rights is typically thought to be a problem plaguing socialism. In contrast, adverse selection presents problems for private ownership because property rights give people "vested interests, which can make it more difficult to motivate them to share their private information with each other" (Myerson (2009), p. 66).

However, according to Myerson, the choice between moral hazard and adverse selection is not a symmetric one. Rather, moral hazard is the fundamental problem threatening economic efficiency. To avoid it, even socialism must give some property rights to individuals, because they are better motivated to work hard in order to maintain their property. But this will give rise to adverse selection problems after all and it therefore delimits the possible advantages of

socialism. So Myerson turns the analysis into an argument against socialism.

Does this analysis show that capitalism is superior to socialism? This is doubtful. While an answer to this question would require a longer discussion, I shall point out two limitations of the analysis here. First, Myerson does not provide full-fledged models of capitalism and socialism, but only models of the incentive problems that arise in the allocation of capital under both systems. Of course, the allocation of capital may be an essential part of evaluating economic systems (Myerson follows von Mises (1935) in this assumption). Nevertheless, deciding the planning controversy might require including in the analysis other markets than the capital market alone. Second, if the analysis is correct, it shows that the choice between capitalist and socialist allocation of capital is subject to a trade-off between different kinds of incentive problems. It does not show that capitalism is superior to socialism in every respect. It may be plausible that moral hazard is a more fundamental problem than adverse selection. However, this is a substantive assumption Myerson makes, which does not follow from the models themselves.

To conclude this section, we note two factual limitations of mechanism design in the search for a better economic system. First, even though mechanism design regards the structure of the economic system as an unknown, as Hurwicz reminds us, the theory did not offer a "third way" besides capitalism and socialism. It is typically thought that the collapse of most socialist economies decided the planning controversy in favour of capitalism. Moreover, it is sometimes lamented that, since capitalism lost its main rival, we lack a grand, alternative idea to capitalism. Mechanism design remained largely silent on such an idea. The theory may be better equipped to exclude systems, rather than finding a superior one to the existing ones, even though some attempts at this have been made. Recently, Eric Posner and Glen Weyl put forward the system of "radical markets", which aims to provide such a third way, integrating features of both capitalism and socialism (Posner and Weyl (2018)).

We will come back to their proposal in the last chapter.

Second, even though Hurwicz draws attention to the potential practical importance of this research programme in designing a "superior" system to the existing one, it did not directly lead to real-world social reforms. This may in part be due to the large-scale focus on economic systems. Models of economic systems are hard to test in practice. Moreover, they neglect the details of economic interactions, such as the rules for exchanges on a specific marketplace. But these details may affect the success of large-scale reforms. This is because changing the fundamental institutions may also change those interactions, for example, by affecting the choices available to individuals. Understanding the details of specific interactions may be a precondition for successful large-scale design.

1.2 The Economist as Engineer

Since the 1980s the main focus of economic design has shifted from grand economic systems to particular markets and marketplaces, and the recommended designs are increasingly implemented. To emphasise their focus on practical problem-solving, Alvin Roth (2002) calls economists engaged in this field *engineers*; Esther Duflo (2017) describes them as *plumbers*.

Various reasons for this shift can be identified. One reason is that considerable progress had been made on game-theoretic models that concern specific transactions. Importantly, these included auctions, a field of research initiated by William Vickrey (1961), and matching under preferences. For an illustration of the latter, consider David Gale and Lloyd Shapley's "marriage market":

> "A certain community consists of n men and n women. Each person ranks those of the opposite sex in accordance with his or her preferences for a marriage partner. We seek a satisfactory

> way of marrying off all members of the community... we call a set of marriages *unstable* (and here the suitability of the term is quite clear) if under it there are a man and a woman who are not married to each other but prefer each other to their actual mates."
> Gale and Shapley (1962), p. 388

They formalise the marriage market and they define a simple procedure in which men iteratively propose to women, who reject all but their most preferred proposal in each step, which always produces stable matchings. The way in which this abstract formulation of a matching problem finally got applied to the real world provides an interesting case study in economic engineering. Roth (1984) considers the matching market that allows medical graduates in the US to find training positions in public hospitals. This market had been subject to severe market failures before a matching procedure had been introduced. This procedure can be shown to be equivalent to Gale and Shapley's mechanism, and it temporarily cured the market failures. However, Roth also shows that the market was failing again when increasing numbers of married couples entered the market. Adding couples to the model, it can be shown that stable matchings may not exist, which provided an explanation for the new market failure. Roth was later commissioned with the redesign of this market, which led to one of the great success stories of market design that was a key motivation behind the 2012 Nobel Prize to Roth and Shapley.

Matching theory has become an important tool for economic engineers and its results have been applied to various labour markets, school choice, or for matching organ donors with transplant patients (see Roth (2018) for an overview). While the design of auctions – in particular for radio spectra – has been subject of detailed methodological analyses in the philosophy of science (e.g. Guala (2005) chapter 8, Alexandrova (2006)), matching has not been analysed in detail before. I address this gap in the literature: in chapter 3, I give a detailed methodological analysis of the redesign of the matching market

for medical graduates. Matching markets will be a recurrent theme in this thesis: in chapters 4 and 5, we will consider ethical issues that specific matching problems raise, namely matching refugees with host countries and kidney donors with patients, respectively.

A further reason for the development of economic engineering is the rise of experimental economics. Engineers and plumbers are aware that minor changes in a marketplace can cause it to unravel in ways that simple models often cannot predict. Therefore, experimental testbeds increasingly complement mathematical models in the development of well-working market rules (see Guala (2005) for a methodology of experimental economics). Besides laboratory experiments, market designers make use of natural experiments, statistical analyses, simulations, etc. It is also common for economists who are (re)designing a particular market to draw upon the expertise of participants in that market, as when medical practitioners are consulted in the design of a kidney exchange.

The methodology of economic engineering is relevant to important debates in the philosophy of science. In particular, there has been extensive debate about how (or whether) economic models create knowledge about the social world.[7] I add to this debate in chapter 3 by giving a general epistemology of economic engineering. According to my account, economic theory encodes counterfactual information in its models about how interventions in a market would change the market outcomes, and we can learn about this counterfactual information by manipulating the models. These manipulations may include de-idealisation techniques, as when couples are added to a simple matching model; and the complementary use of empirical methods. This addresses what has been labelled a "paradox" in the philosophy of economics, viz. that models

[7]The following is an (incomplete) list of contributions to this big debate: Alexandrova (2008); Cartwright (2009); Guala (2005); Hausman (1992); Kuorikoski and Lehtinen (2009); Mäki (2009, 2017); Morgan (2001); Reiss (2012); Rubinstein (2006, 2012); Schelling (2010); Sugden (2000). Alexandrova and Northcott (2013) and Alexandrova and Northcott (2015) take a sceptical stance, calling into doubt that economic models yield explanations.

provide insights into the social world even though they are highly idealised and incorporate false assumptions.

Following success stories such as the one above, there is increasing demand from policy-makers for economists "repairing" deficient markets, or creating new markets. What kinds of markets do economic engineers recommend? While it is very sensitive to the context what design could yield a successful market institution, a few general constraints on successful markets can be identified.

Incentive compatibility still figures as an important constraint. However, incentive compatibility is defined against the stringent rationality assumptions prevalent in mechanism design, which should in some contexts be relaxed. For instance, Shengwu Li (2017b) defines "obviously strategy-proof mechanisms", which provide cognitively limited agents with weakly dominant strategies in a way that they can recognise these strategies as weakly dominant and thus manage to play equilibria. Obviously strategy-proof mechanisms exclude situations in which game theorists predict equilibrium outcomes, which do not materialise in the real world because the individuals' strategies are too complicated.

In addition to incentive constraints, Roth identifies a variety of constraints on well-functioning markets (see Roth (2013, 2015a)). For instance, markets should be *thick*, that is, they should attract a sufficient number of potential buyers and sellers to facilitate satisfactory trades. However, as markets get thick, they might *congest*: a congested market makes it complicated or time-consuming for traders to identify favourable trades. Furthermore, they should avoid *repugnance*. A repugnant market is one that people do not want individuals participating, even though they do not have externalities. We will come back to repugnance in the discussion of ethics and economic design below.

A lack of any of these features can cause a market to unravel. But there is no one-size-fits-all mechanism that could achieve these features for all kinds

of transactions. Obviously, while few would object to selling one's old bike on ebay, offering one's kidney will arouse considerable repugnance (and is in fact illegal). Likewise, what a good design is depends on the goods to be exchanged in the market in question, its structure, size, etc. Some markets work best when unregulated, others are "crippled by inconsistencies in information, control, incentives, and behavior, and require social management" (McFadden (2009), p. 78). These insights from market design caution against overly ideological debates concerning free markets vs. regulation. Market designers' mixed recommendations that depend on the details of the marketplace suggest that this is a false dichotomy. This might be a reason why, as economic design became more practical, large-scale debates about economic systems increasingly slipped out of focus.

To take stock, we have identified two grand traditions within economic design. The theory of mechanism design provides a framework for comparing and evaluating economic systems. It promised a third way besides capitalism and socialism, but despite its reformist spirit, its contributions remained mainly theoretical. The other is economic engineering, which is compartmentalised and focused on the details of small-scale problems. It is closer to policy-making, but this practical advantage comes at the expense of losing the role of evaluating fundamental economic institutions. We will take up the relation between these two traditions, as well as possible future developments, in the last chapter.

1.3 Economic Design and Ethics

The choice between different mechanisms depends not only on figuring out incentives and efficiency, but it often involves substantive ethical questions. This raises questions about the relation between economic design and ethics. Is there a 'right' ethical theory for economic design? And if so, is preference

utilitarianism this theory, which is assumed as default by many economists outside economic design? If this is the case then economic design should appear as an adversary to many strains of moral philosophy, namely all those opposed to preference utilitarianism. Debates about the moral limits of markets might suggest that this is indeed the case. Many economic designers hold liberal views about their reach. In contrast, some moral philosophers argue for a narrower domain, adducing normative reasons for this view (Sandel (2012)).[8]

However, this is an oversimplified and indeed a false picture of how economic designers treat ethics. They have suggested a multitude of properties, and mechanisms implementing them, that do not speak to (preference) utilitarianism. Examples are procedural fairness (Klaus and Klijn (2009)), or the elimination of justified envy (Abdulkadiroğlu and Sönmez (2003)), to only mention two properties that can be implemented by matching mechanisms. Given this multitude of properties and associated mechanisms, the crucial question is, how should the mechanisms be chosen that ought to govern a specific interaction? Moreover, who should choose it – in particular, what is the right division of labour between economic designers, ethicists and policy-makers?

Li (2017a) provides a simple and flexible framework for thinking about the relation between economic design and ethics. Suppose we wish to design an institution, say a market, that will govern some prespecified interactions. There is a set of feasible designs D and a set of consequences C. A function $f : D \longrightarrow C$ determines what consequences the feasible designs would lead to. To each of the consequences corresponds a value judgement J, specified by a value function $v : C \longrightarrow J$. Defining $v(d) := v(f(d))$ for all $d \in D$, we extend the value function to range over the designs themselves.

This model captures only consequentialist ethical theories. Redefining the domain of the value function to cover design-consequence pairs $v : D \times C \longrightarrow J$ allows including non-consequentialist theories, in which the value may change

[8] See Besley (2013) for an interesting discussion of how moral philosophy and economics treat the moral limits of markets.

depending on the mechanism in use.

Now, it could be thought that there is a stark division of labour: market designers investigate f, whereas ethicists and policy-makers investigate v, and there is no overlap between their respective tasks. However, this is false. To see this, note that, how f is specified partly depends on v. Market designers do not specify all consequences of a design, because the set of consequences may be large and many properties of consequences are relatively uninformative. Instead, they classify consequences along properties that are morally relevant. This requires a theory of what properties are morally relevant in the first place and it requires an interpretation of those properties within the model in question. For instance, remember Gale and Shapley's marriage model. Here, f could include statements such as, "deferred acceptance procedures lead to fair matchings". This presupposes that fairness is a relevant property. In addition, it requires an interpretation of fairness within this model, as, for instance, fairness in a monetary market will have a different meaning than fairness in a matching market.

For these reasons, the model should be extended again to include a set of "intermediate judgements" I that state what ethically relevant properties a design-consequence pair implements, such as efficiency, fairness, transparency, etc. With each design-consequence pair is associated an intermediate judgement, so there is a function $g : D \times C \longrightarrow I$. This function states things like, "d implements fair matchings". With each intermediate judgement is associated an overall judgement, $v : I \longrightarrow J$. This function states things like, "all things considered, we should implement d".

Li then proposes the following division of labour: economic designers study g; ethicists and policy-makers study v. This division of labour means that "the theory and practice of [economic] design should maintain an informed neutrality between reasonable ethical positions" (p. 717). Economic designers should be informed in order to define all the relevant elements of I in a way

that is consistent with their standard meaning in ethics. They should be neutral, that is, leave it to ethicists and policy-makers to resolve fundamental ethical disagreements about v.

It should be added to Li's analysis that his proposed division of labour also requires of moral philosophers to know some results from economic design. Obviously, they need knowledge about I, which is the domain of the value function v. They must know the meanings of these intermediate judgements: if an intermediate judgement asserts that "d produces fair matchings", they should know how fairness is defined in this context. Moreover, they should know what the relevant elements of I are, that is, what combinations of features are feasible in a given market. For instance, it is not a reasonable position to demand fairness and efficiency if this demand is impossible to meet. (We will grapple with this problem concerning the distribution of refugees in chapter 4.)

The second part of the thesis discusses v with regard to two matching problems. In chapter 4, I discuss some ethical issues concerning the distribution of refugees over host countries. In chapter 5, I give an argument for kidney exchange. The two chapters differ in how coarse-grained they partition I. Concerning the distribution of refugees, I apply basic matching models and mechanisms, which provide insights into some of the ethical trade-offs that arise in this context. In contrast, concerning kidney donations, I do not consider specific matching procedures. Rather, the motivation is that some countries ban the possibility of matching donors and recipients altogether, by restricting live organ donations to close relatives who may be incompatible. I argue for the implementation of matching markets in this context, because these markets are instrumental in meeting a moral obligation, namely to donate effectively.

Does the practice of economic design generally follow Li's ideal of an informed neutrality concerning ethical theories? It could fail neutrality or informed-

ness. Concerning neutrality, consider how economic designers treat repugnance. Roth (2007) introduces repugnance as a factual constraint on markets. For instance, if there are many people who object fiercely to a monetary market in kidneys, such a market might unravel. However, Roth has also advocated removing financial disincentives for donating kidneys, or designing a market in which government buys and allocates kidneys.[9] So repugnance is in this case not only seen as a factual constraint on markets, but as one that ought to be removed.

Concerning informedness, if economic designers are more aware of some ethical theories than others, this can lead to a biased provision of intermediate judgements. For instance, given the prevalence of preference utilitarianism in the broader discipline of economics, it is plausible that the intermediate judgements are predominantly consequentialist.[10]

Failing neutrality is not so problematic if economic designers do not fail informedness. It seems fine (and perhaps inevitable) that an economic designer would make all-things-considered judgements (statements about v) if the function $g : D \times C \longrightarrow I$ is defined in the following way: first, I contains elements from a large class of ethical theories, as opposed to, say, elements corresponding to consequentialist judgements but no elements corresponding to non-consequentialist judgements; and second, I is partitioned in a way that respects the standard meaning of ethical concepts. A moral philosopher, or a policy-maker, would then be free to disagree with the all-things-considered judgement by this specific economic designer. To a moral philosopher who disagrees in this way, economic design would not, as a discipline, appear as

[9] See `https://www.theatlantic.com/business/archive/2015/10/give-a-kidney-get-a-check/412609/` for removing financial disincentives. For government buying and allocating kidneys, see the session summary of the 2016 Normative Ethics and Welfare Economics Conference, Harvard Business School, `https://www.hbs.edu/faculty/conferences/2016-newe/Documents/Session%20IV.pdf`. Both accessed on 18/04/2019. I should add that Roth does not make these recommendations in his academic articles.

[10] It is also easier in practice (e.g. more parsimonious with regard to notation) to evaluate consequences, rather than design-consequence pairs, which might additionally favour consequentialist theories.

an adversary, even though the particular economic designer with whom she disagrees might. An informed debate would be possible. In contrast, if the elements of I were defined and partitioned in an uninformed way – important elements are missing, the partition is skewed – then a moral philosopher studying v could perceive the discipline of economic design as an adversary, even if economic designers were silent about all-things-considered judgements (they only studied g). Moreover, if moral philosophers lack the technical expertise for redefining the set of intermediate judgements and a corresponding g' in a less problematic way, an informed debate is not possible.

Since informedness is key, this argument provides a rationale for the importance of ethical theory for economic design. In particular, it may suggest that ethics should form an integral part of the curriculum of economic designers.

1.4 Conclusion

Economic design is at the intersection between positive and normative economics. Its origins lie in the debates over the relative merits of centrally planned versus market economies. Mechanism design provides important insights into these debates by examining the incentives that economic systems provide people with. Subsequently, economic design became focussed on specific transactions, such as auctions, or matching problems. Economic designers have developed an engineering approach that seeks to devise well-functioning, real-world institutions, but the large-scale system design increasingly slipped out of focus. The choice of a mechanism often involves ethical considerations and trade-offs. Arguably, economic designers need not always live up to the ideal of informed neutrality concerning reasonable ethical theories: they need not always be neutral, but they should be literate in ethics. This provides a rationale for the importance of ethics in economic theory.

Economic design raises a variety of philosophical and ethical issues: how do

rational individuals interact with each other through institutions? And what does this tell us about how our economic system should be organised? How can economic models be used for predicting how market outcomes will change as we intervene in markets? How ought we to design specific markets and other institutions? In the next chapters, I hope to provide some insights into these questions.

1.5 Overview of the Chapters

In the next chapter, I examine the implications of incentive compatibility for notions of rationality. Against the orthodoxy to view Nash equilibrium as "the embodiment of the idea that economic agents are rational" (Aumann (1985), p. 43), some theorists have proposed non-classical notions of rationality in games, asserting the possibility of rational choices that do not constitute equilibria. I discuss an implication that non-classical theories, such as constrained maximization, or team reasoning share: they are at odds with the requirement that institutions be designed to be incentive-compatible, that is, that they implement desired social goals in equilibrium. Proponents of non-classical theories face a choice between three options to resolve this conflict: either they deny that incentive compatibility is required as a constraint on institutional design, or they deny that individuals interacting through the designed institutions are rational, or they accept that their theories do not apply to institutional design. I discuss these options critically and I argue that institutional design presents a challenge to non-classical theories of rationality to which proponents of those theories haven't responded convincingly.

Chapter 3 gives an account of the epistemology of economic engineering. I introduce a case study – the reform of a matching market for medical residents – which hasn't been previously studied in the philosophy of science and which challenges some views on economic engineering. In particular, some philoso-

phers of science have been critical of the contribution of economic theory to economic engineering. However, in our case study, the use of economic models is key for understanding and redesigning the market. I suggest a more general account according to which models, e.g. from game theory, allow formulating policy goals. Complemented by experiments and other methods, they project institutions implementing those goals. Directed graphs are introduced to illustrate how in this process, economic institutions are treated as variables to be intervened on. Finally, I argue that the creation of knowledge in economic engineering urges caution on a recent call in the philosophy of science for more aggressively evaluating the use of models in economics.

The second part of the thesis examines ethical questions that the design of specific institutions raises. Chapter 4 is about the distribution of refugees over host countries. It has been suggested that their distribution can be made more fair or efficient if policy makers take into account not only numbers of refugees to be distributed but also the goodness of the matches between refugees and their possible host countries. There are different ways to design distribution mechanisms that incorporate this practice, which opens up a space for normative considerations. In particular, if the mechanism takes countries' or refugees' preferences into account, there may be trade-offs between satisfying their preferences and the number of refugees distributed. I argue that in such cases, it is not a reasonable policy to satisfy preferences. Moreover, conditions are given which, if satisfied, prevent the trade-off from occurring. Finally, it is argued that countries should not express preferences over refugees, but rather that priorities for refugees should be imposed, and that fairness beats efficiency in the context of distributing asylum. The framework of matching theory is used to make the arguments precise, but the results are general and relevant for other distribution mechanisms such as the relocations currently in effect in the European Union.

Chapter 5 discusses the ethics of kidney exchange. The best treatment for

end-stage renal disease is the transplantation of a live donor kidney, but many people cannot donate to their loved ones because they are incompatible. Kidney exchange promises relief. Kidney exchange programmes use centralised procedures to match donors with recipients in a way that maximises the quantity and quality of transplants. However, kidney exchange has met ethical concerns, and the transplant laws in many countries render it impossible. I give two arguments for the implementation of kidney exchange programmes. The first is that they are instrumental in meeting a moral obligation, namely to donate effectively. The second is that they may increase the motivation for altruistic donations, because the donation of one kidney may trigger > 1 life savings. Moreover, ethical concerns are considered that are embodied in transplant laws preventing the implementation of kidney exchange, and it is argued that they can be overcome.

The final chapter identifies open questions for future research and it closes with some thoughts about the future trajectory of economic design.

Chapter 2

Of Rats and Rationality

Institutional Design and Rationality in Games

Michael Vann describes an episode of French colonialism in the city of Hanoi. Threatened by a rat infestation that was apprehended to cause an outbreak of the pest, the colonial administration declared that it would pay a bounty to Vietnamese residents for each rat-tail that they would bring them. The Vietnamese handed in thousands of tails, yet the rat population did not diminish. The reason became evident when officials discovered that the "exterminators" were cutting off rat-tails while leaving the rats alive and able to breed and produce more valuable tails. Later, the authorities also detected rat farming in the suburbs of Hanoi. "One can only imagine the frustration of the municipal authorities, who realized that their best efforts at *dératisation* had actually increased the rodent population by indirectly encouraging rat-farming. Evidently, this was not what the French had in mind when they encouraged capitalist development and the entrepreneurial spirit in Vietnam" (Vann (2003), p. 198, emphasis in the original).

The colonial rulers were apparently not aware of what economists now almost universally acknowledge: institutions – such as markets for disinfestations – should be designed to be *incentive-compatible*, that is, they should not bring about adverse incentives that produce unintended outcomes. Game theory plays a central role in the design of incentive-compatible institutions: institutional designers devise games that implement some desired social goals in equilibrium, and they aim to make institutions resemble those games. This methodology is based on a specific notion of *rationality*, as well as on a rationality assumption. The notion of rationality is, roughly, that rational agents play equilibria, and the rationality assumption is that the people who will interact through the designed institution are rational. Together, these assumptions imply that the designed institution will bring about the desired social goals. This methodology coheres with the logic of incentives, because in equilibrium, individuals have no incentives to deviate from their actions.

However, some theorists maintain that there are rational choices that do not constitute equilibria. Such "non-classical" theories include team reasoning and constrained maximization, both of which will be introduced in the next section. This chapter sheds light on the contrast between classical and non-classical theories from the perspective of institutional design. It shows that non-classical accounts are at odds with the constraint of incentive compatibility: what institutional designers expect individuals to do when they impose this constraint is in many cases inconsistent with what non-classical rationality requires them to do. The chapter then identifies the options that non-classical theorists can choose from in order to resolve this conflict: either they reject incentive compatibility as a constraint on institutional design; or they deny that the individuals interacting through the designed institutions are rational; or they concede that their theories do not apply to institutional design.

I shall critically discuss these options in turn. First, rejecting the constraint of incentive compatibility risks diminishing social welfare and empirical evidence

suggests that it would indeed diminish social welfare. Second, treating people as irrational has been unpopular among non-classical theorists themselves and it shows an advantage of classical relative to non-classical theories. Third, conceding that non-classical theories do not apply to institutional design seems to present an argument against those theories. More charitably, it could be seen as delimiting the scope of applicability of those theories. But non-classical theorists should then give an account of the conditions under which their theories supposedly apply. However, a convincing account is to date owed and it is difficult to see how such an account could be provided. I conclude that institutional design constitutes a challenge to non-classical theories of rationality to which proponents of those theories haven't given a convincing response.

If a theorist argues for a contextual account of rationality (that is, opts for the third option above), there is an interesting implication: the meaning of "rational" may depend on the concept of institution, because institutional contexts delimit the possible scope of non-classical theories of rationality. For this reason, a sideline of this chapter is to offer some insights into the concept of institution. In particular, I shall make some constructive criticisms of a recent theory of institutions by Francesco Guala and Frank Hindriks. According to their theory, institutions are "rules-in-equilibrium". I argue that their account provides valuable insights by emphasising that successful institutions implement desirable outcomes in equilibrium. I suggest two amendments to this account: first, they should include mechanisms (in a sense that will be defined) as essential parts of institutions. Second, their account takes incentive compatibility for granted, which excludes entities typically thought of as institutions, and it may obscure the fact that institutions can succeed or fail. From the perspective of institutional design, it is preferable to regard institutions as entities governed by mechanisms that may or may not be in equilibrium.

This chapter is organised as follows. The next section presents classical and

non-classical theories of rationality in games and Section 2.2 introduces the concept of incentive-compatible institutional design. Section 2.3 brings together the previous two sections: I show that non-classical theories of rationality are at odds with incentive compatibility and I discuss the options that non-classical theorists have to resolve this conflict. Section 2.4 picks up the discussion about the concept of institution, and Section 2.5 concludes.

2.1 Notions of Rationality in Games

How do rational individuals interact with each other? This question has occupied many philosophers and economists alike. Their responses can be classified into two distinct families, which I call "classical" and "non-classical", respectively. Each family subsumes various theories of rationality in games. I shall first introduce classical theories and give their defining characteristic, which is that they place Nash equilibrium at centre stage. Then I present two prominent examples of non-classical theories, namely constrained maximization and team reasoning.

In order to introduce these notions of rationality, it is convenient to consider some simple games in normal form. For instance, consider the Hi-Lo game in Figure 2.1. In this game, players I and II simultaneously choose Hi or Lo. Each combination of their choices results in a cell, specifying both players' payoffs. Player I's payoff is shown on the bottom left and player II's payoff on the top right of each cell. The payoff numbers refer to the players' utilities. These summarise everything that is motivationally relevant to the players (I shall have more to say about this below). We make the standard assumption that players' utilities are measured on an interval scale, that is, the point of zero utility and the units of utility are arbitrary and the ratios of the differences between the utilities are non-arbitrary. Interpersonal comparability of players' utilities is not assumed. Moreover, the utilities are not transferable between

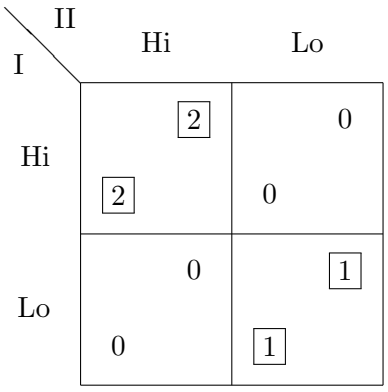

Figure 2.1: Example of a Hi-Lo game. Player I's payoff is shown on the bottom left and player II's payoff on the top right of each cell. A square around a payoff number denotes a player's best reply to a possible choice of the opponent.

the players.

Under these assumptions, what should player I choose? This obviously depends on what player II chooses: if II chooses Hi, I's best reply is to play Hi too, if II chooses Lo, I's best reply is to play Lo. Because of the symmetry of the game, the same holds for player II. A player's best replies to the opponent's possible choices are marked with squares around her payoff numbers in the figure. There are two outcomes in the game in which the players' actions are best replies to each other: both players play Hi, or both players play Lo. Furthermore, instead of choosing an action with certainty, the players may come up with a strategy to choose both available actions with non-zero probabilities. For example, they could choose their actions depending on the outcome of the throw of dice. If players' strategies include such plans of action in which they mix their choices, then there is a third outcome in which the players' strategies are best responses to each other: both players play Hi with probability 1/3 and Lo with probability 2/3, which yields both players an expected payoff of 2/3.

In general, a strategy profile in which each player's strategy is a best reply to all

the other players' strategies (where strategies may be pure or mixed), is a *Nash equilibrium*. In the following, I will drop "Nash" when it is clear that Nash equilibrium is at stake, and I will call individual strategies that are played with a positive probability in some equilibrium, "equilibrium strategies". According to classical theories of rationality, then, a strategy is rational only if it is an equilibrium strategy. Because in an equilibrium, no player has incentives to deviate from her strategy, classical theories establish tight links between rationality and incentives. This leads Robert Aumann to state the classical view thus:

> "The Nash equilibrium is the embodiment of the idea that economic agents are rational; that they simultaneously act to maximize their utility. If there is any idea that can be considered *the* driving force of economic theory, that is it. Thus in a sense, Nash equilibrium embodies the most important and fundamental idea of economics, that people act in accordance with their incentives." (Aumann (1985), p. 43)

According to the most common classical theory, equilibrium strategies are not only necessary, but also sufficient for rationality (e.g. Binmore (2007)). This theory is sometimes challenged because games can have many equilibria, some of which may be less than optimal. For instance, in the Hi-Lo game, despite there being three equilibria, it seems "trivial" to many that (Hi, Hi) should be the unique rational outcome of this game (e.g. Gold and Sugden (2007), p. 284), and "paradoxical" that standard game theory does not solve for this outcome alone (e.g. Bacharach (2006), p. 44 et seq., Guala (2018b)).

However, even though this critique is sometimes supposed to motivate a departure from classical theories (e.g. Gold and Sugden (2007), p. 284 et seq.), it should be noted that refinements of Nash equilibrium, according to which playing equilibrium strategies is necessary, but not sufficient for rationality,

can rule out "bad" equilibria. John Harsanyi and Reinhard Selten developed a general theory that selects a unique equilibrium in a large class of games (including all games in normal form), which is the rational outcome, according to their theory (Harsanyi and Selten (1988)). For instance, in Hi-Lo games, their concept of "payoff dominance" implies that only the Hi strategy is rational to play, thus ruling out the two dominated equilibria.[1]

In contrast to classical theories, according to which equilibrium strategies are a necessary condition for rational play (and in some of which equilibrium strategies are also a sufficient condition for rational play), non-classical theories of rationality in games maintain that equilibrium strategies are neither necessary nor sufficient for rational play. The motivation for proposing these theories is typically an unease with inefficient equilibria, and so the theories imply that rational players can improve upon inefficient equilibria. Apart from Hi-Lo games, where this unease may fall short of necessitating non-classical theories (because some classical theories accommodate this concern), Prisoners' Dilemma (PD) games are typically adduced as evidence for the alleged shortcomings of all classical theories. An example of a PD game is shown in Figure 2.2. In this game, it is a dominant strategy for both players to defect, therefore (Defect, Defect) is the unique, dominant-strategy equilibrium. But (Cooperate, Cooperate) strictly Pareto-dominates this equilibrium, that is, were the players able to achieve this outcome, both would be better off. According to classical theories, they could never achieve this outcome since both players would have incentives to deviate by playing Defect. Instead of interpreting these incentives as a constraint on what can rationally be achieved, proponents of non-classical theories consider this a weakness of the orthodoxy. We next consider two prominent examples, constrained maximization and team reasoning, in more depth.

[1] However, Bacharach (2006) criticises Harsanyi and Selten's theory on the basis that it is aimed at equilibrium selection, that is, determining a unique rational outcome. According to Bacharach, a theory of rationality should reflect intuitions of rationality, but these are indeterminate in many other games (see p. 60 et seqq.).

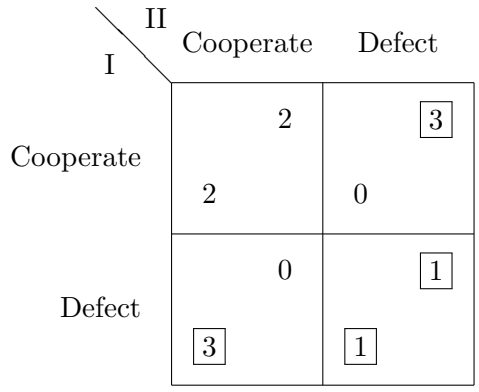

Figure 2.2: Example of a Prisoners' Dilemma game.

Constrained maximization. David Gauthier is a prominent defender of the view that rational players can improve upon suboptimal equilibria in situations in which there is ground for mutual trust. He locates rationality at the level of agents' dispositions to choose (rather than at the level of strategies for choices), and he defines "constrained maximization" as the disposition to choose cooperatively if the other agent(s) have the identical disposition, and non-cooperatively otherwise.[2] A population of constrained maximizers playing PDs could therefore improve upon the non-cooperative outcome that "straight maximizers" achieve, who simply choose best replies.

This view faces the problem that straight maximizers could simply exploit constrained maximizers in PD-situations. The latter would then be worse off as a consequence of their disposition to cooperate, which would constitute an odd account of practical rationality. In order to exclude this possibility, Gauthier introduces the condition of "translucency", according to which an agent's disposition for constrained or straight maximization is known to others in the population with a positive probability. Thus, when agents within the population play against each other, there is a probability that constrained

[2]The canonical exposition of his theory of constrained maximization is in chapter 6 of Gauthier (1987). His most recent attempt at rationalising cooperation in PD games is Gauthier (2015). For a critique of the latter, see van Basshuysen (2016).

maximizers will recognise each other and therefore cooperate, and there is a probability that a constrained maximizer will fail to recognise a straight maximizer in the population and will therefore be exploited. According to Gauthier, it is rational for an individual to choose the disposition to constrained maximization if this maximises her expected utility. This is the case, roughly, when the choice of disposition is sufficiently translucent and the proportion of constrained maximizers in the population sufficiently high.[3]

Thus, constrained maximization is designed as a strategy to rationalise cooperation in PDs by placing rationality at the level of dispositions and requiring that these dispositions be translucent. More generally, the same argument supposedly rationalises non-equilibrium play in games in which an outcome strictly Pareto-dominates all equilibria. Moreover, in games, such as Hi-Lo, in which there is a Pareto efficient equilibrium, the theory solves for this equilibrium. So in short, for Gauthier, Pareto efficiency takes the place of equilibrium as the criterion for rationality.

Team reasoning. A number of theorists have criticised concepts of rationality that derive from best-reply reasoning as too individualistic. These theorists argue that players may sometimes reason from the perspective of "we", instead of "I", that is, as a team.[4] When they reason as a team, players are assumed to identify the action profiles (instead of their individual actions only) that best promote the common interests of the team they form part of. Then they

[3]For constrained maximization to maximise utility requires a combination of the two conditions – level of translucency and proportion of constrained maximizers in the population – such that the more constrained maximizers there are, the higher the risks can be that constrained maximizers fail to cooperate and be exploited by straight maximizers. For a detailed exposition, cf. Gauthier (1987), p. 176 et seq.

[4]The following are some of the main contributions to team reasoning in games. Robert Sugden introduced team reasoning to game theory in Sugden (1993). Bacharach (2006), which was completed by Natalie Gold and Sugden after Michael Bacharach passed away in 2002, has the status of a classic. Some recent developments are Sugden (2011, 2015) and Karpus and Radzvilas (2018). For comparisons of different theories of team reasoning, see Gold and Sugden (2007) and Karpus and Gold (2017). I shall present what I take to be the core tenets of team reasoning here, rather than the details over which some of the theorists are at variance with each other.

choose their individual actions that jointly generate those action profiles. For example, if the players of a PD game form a team, then, assuming that universal cooperation is identified as the unique optimal outcome for this team, the players will choose to cooperate.

For individuals to reason as a team requires an identification of these individuals with the team that they jointly constitute. This identification can also be described as a transformation from individual to collective agency (Gold and Sugden (2007), p. 292). As a result of this transformation, the players put aside their individual interests and act upon the interests of the team. This process raises two questions: how do individual interests convert to team interests? And why would rational players act upon the team interests, which might (depending on the answer to the first question) require them to sacrifice their individual utility to benefit the team? We shall consider these questions and some proposed answers in turn.

According to some early versions of the theory, as a consequence of the described agency transformation individual utilities map into 'team utilities', the maximisation of which is taken to best promote the interests of the team. For instance, if the team utilities in a given game equal the average of the agents' individual utilities, then advancing the team interests would amount to maximising the average of the agents' individual payoffs. However, in many games, this would benefit some players at the expense of others in ways that may seem unjust. For example, suppose in the example of the Prisoners' Dilemma game in Figure 2.2, the individual utilities from defecting if the other player cooperates were 5 instead of 3, while the game is otherwise identical to before. In this game, the asymmetric outcomes (Defect, Cooperate) and (Cooperate, Defect) maximise the average of the agents' utilities. But these outcomes may seem to demand an undue sacrifice from the cooperating member to benefit the non-cooperating member of the team (remember that utility is non-transferable

so a compensation is not possible).[5] Sugden (2011, 2015) proposes a more promising approach: team play should yield the players a *mutual advantage*, requiring that the outcome is at least as good as the players' maximin payoff (that is, the utility that each player can achieve independently of the other players). Karpus and Radzvilas (2018) give a possible formal characterization of mutual advantage in normal form games. They present a measure that can be applied to calculate which outcome(s) in a normal form game maximise the players' mutual advantage, relative to possible reference points. The outcome(s) that maximally advance team interests are then defined as those that yield maximum mutual advantage.

If rational players could reach action profiles that maximise mutual advantage, or some other measure of team interest, they could thereby implement outcomes which in many games yield better individual payoffs to the players than equilibrium outcomes. But this would often require them to choose contrary to their incentives, for instance, when cooperation is identified as the outcome that best advances the team interest in a PD game. Why would rational players do this? According to proponents of team reasoning, as a consequence of the agency transformation, the joint action of the team can be described as rational. The rationality of this joint action implicates the rationality of the individual choices of the team members, which constitute the joint action.[6] There is disagreement between different proponents about whether or not the agency transformation is itself a requirement of rationality. For instance, Hur-

[5] Moreover, averaging players' utilities is meaningful only if their utilities are interpersonally comparable, which is an assumption that is not typically licensed in game theory (see above), which may be seen as a further problem for these mappings.

[6] I should add that team reasoning is sometimes put forward as a descriptive theory of interactive choice, rather than as a theory of rationality. For instance, when Sugden states his own position in Bacharach (2006), he claims to be "less concerned with the validity of team reasoning, treating it only as an idealised model of a form of reasoning which people in fact use, whether justifiably or not" (p. xxii). However, this does not keep him from suggesting in various other places that according to team reasoning, "the rationality of each individual's action derives from the rationality of the joint action of the team" (Sugden (2003), p. 167; also cf. Gold and Sugden (2007), p. 285). Therefore, it could be doubted that he sees team reasoning merely as a descriptive theory. Be that as it may, since our concern here is rationality, we are interested in team reasoning insofar as it is proposed as a normative theory.

ley (2005a,b) argues that team identification is the result of rational choice. Other theorists hold that team identification, or failure thereof is not the result of conscious deliberation and choice, but rather of the workings of psychological mechanisms that are in effect when certain conditions are satisfied (e.g. Bacharach (2006)). (See Karpus and Gold (2017) for a comparison of different proposals.) We will come back to these conditions below, but the details of these disagreements are insignificant for present purposes. It suffices that, according to theories of team reasoning, there are conditions under which individuals identify as a team, and when they are satisfied, rationality may require those individuals to play non-equilibrium strategies.

To sum up, we have distinguished classical from non-classical theories of rationality in games. The distinguishing feature is that the former require of rational strategies to be equilibrium strategies, whereas the latter seek to rationalise non-equilibrium strategies that may lead to outcomes that dominate equilibria. Constrained maximization and team reasoning were introduced as examples of non-classical theories.[7] Even though they arrive at their conclusions for different reasons – team reasoning through group identification, constrained maximization through individuals' disposition to choose – they share some of their conclusions. In particular, both imply that there are conditions under which rational agents cooperate in PD games.

To conclude this section, I shall add a clarification regarding non-classical theories: these theories seek to rationalise non-equilibrium strategies while insisting that the players' utilities summarise everything that is motivationally relevant for the players. For instance, suppose in the PD game in Figure 2.2, the payoff numbers refer to dollar values instead of players' utilities. In this game, the players may be motivated by all sorts of considerations. For example, they

[7] These are not the only non-classical theories that have been proposed, but they may be the ones most widely discussed in the literature. Another example of a non-classical theory is hypothetical bargaining.

might choose to cooperate because they sympathise with the opponent. If they cooperate from sympathy, this would not falsify the classical analysis of the PD, but would merely show that the PD game in which the payoffs are dollars should not be analysed as a PD game in which the payoffs are utilities, because doing so would leave out considerations – sympathy – that are motivationally important to the players. Proponents of non-classical theories emphasise that their theories *cannot* be reinterpreted as payoff transformations in games that could then be analysed classically. In the case of team reasoning, an argument to this effect is the following. Take the PD game in Figure 2.2 and suppose for the sake of argument that the team consisting of players I and II prefers the outcome (Cooperate, Cooperate) to (Defect, Defect) and it prefers (Defect, Defect) to the asymmetric outcomes. So if they reason as a team, the players will rationally cooperate. But if we simply substituted the team preferences for the original, individual utilities, we would get an instance of the Hi-Lo game, in which (Cooperate, Cooperate) corresponds to the Hi equilibrium, (Defect, Defect) corresponds to the Lo equilibrium, and there is a third equilibrium in mixed strategies. All three equilibria are rational to play, according to the subset of classical theories that take all equilibrium strategies to be rational. Therefore, the argument concludes, team reasoning cannot possibly be accommodated by applying those classical solution concepts to games that are the result of a payoff transformation – they must involve a deeper agency transformation.[8] To conclude, rather than accommodating factors, such as sympathy, in players' utilities and solving games classically, in many games non-classical theories require of rational agents to act against their incentives.

[8]This argument is from Bacharach (2006), p. 173. For an equivalent line of argument that takes Hi-Lo instead of PD as the initial game, see Karpus and Radzvilas (2018), p. 6 et seq. The latter also mention that, while agency transformations go beyond payoff transformations, they might *involve* payoff transformations. This possibility is not precluded by the above argument. However, their article (and almost all the literature on team reasoning to date) assumes that this does not happen, and we shall follow this assumption here.

2.2 Incentive-Compatible Institutional Design

Institutions can develop 'naturally', but they can also be designed towards specific social goals. In the latter case, a social planner defines the goals that she wants the outcomes of some interactions to possess. A designer then seeks to devise an institution that will govern these interactions in a way that brings about the defined goals. Typically, the designer models the interactions in question as games, using game theory to determine how the interactions would result in different outcomes depending on different "rules of the game", or *mechanisms*, of which she chooses the one that best promotes the defined social goals.[9] Designers emphasise that mechanisms must be *incentive-compatible* in order to reliably produce desirable outcomes. We shall consider a simple example from Roger Myerson to introduce the concept of incentive-compatible mechanism.

Following Myerson (2008), we consider the design of a mechanism when the seller of a single, indivisible item faces one potential buyer. It is commonly known that the seller values the item either at \$0 or at \$80, and that the buyer values the item either at \$100 or at \$20. For simplicity, we assume that traders simply seek to maximise their profits, that is, dollar values define their utilities. So both the buyer and the seller profit from trade unless a type 80 seller (a "strong seller") faces a type 20 buyer (a "strong buyer"). Whether they are weak or strong is their private information, but it is commonly known that each is of a strong or a weak type with an independent probability of $1/2$.

We wish to come up with a mechanism that determines, depending on the

[9] The reader may note an ambiguity: "mechanism" can be used to refer to the institutional rules themselves, or to refer to a formal model of those rules. This ambiguity is prevalent in the mechanism design literature, as Guala (2005) notes (p. 163, footnote 4; also cf. Guala (2001)). Moreover, this ambiguity extends to adjectives characterising mechanisms, in particular, "incentive-compatible". The ambiguity has no problematic implications for my purposes; equivalently but more cumbersomely, we could have restricted the use of "mechanism" to refer to the model of institutional rules, so that an institution would not be governed by a mechanism, but by the rules that the mechanism is a model of. In the next chapter, I discuss the use of models in institutional design in more depth.

	Buyer's value	
	[strong]	[weak]
Seller's value	$20	$100
[strong] $80	0, *	1, $90
[weak] $0	1, $10	1, $50

P(trade), E(price if trade)

Figure 2.3: Split-the-Difference mechanism. From Myerson (2008).

traders' valuations, whether trade should happen and if so, at what price. Since the valuations are the traders' private information, a mediator will ask them for their valuations. Depending on the information provided, she will announce whether the object will be sold and if so, at what price. What could this mechanism look like? A natural idea is the "Split-the-Difference mechanism". It specifies that, whenever the buyer's valuation is higher than the seller's, the item will be sold for a price that equals the difference between their valuations. If both are strong, the item won't be sold. The table in Figure 2.3 shows this mechanism. Each cell corresponds to a combination of the players' types. The first number in a cell denotes the probability that the item will be sold, and the second number the price in cases in which it is sold.

But will the players honestly report their types? This cannot be expected. For example, if the seller is weak (i.e. values the item at $0), she would gain from lying about her type: supposing the buyer reveals his type honestly, the seller's expected profit from revealing weakness is $1/2(10-0)+1/2(50-0)=30$. If she instead claims to be strong, her expected profit is $1/2(90-0)=45$. Thus, it is not an equilibrium of Split-the-Difference that the traders honestly reveal their types: the mechanism is incentive-*in*compatible.

Conversely, a mechanism that, unlike Split-the-Difference, does make it an equilibrium for players to honestly report their types is called incentive-compatible. Continuing the example, we can ask what constraints an incentive-

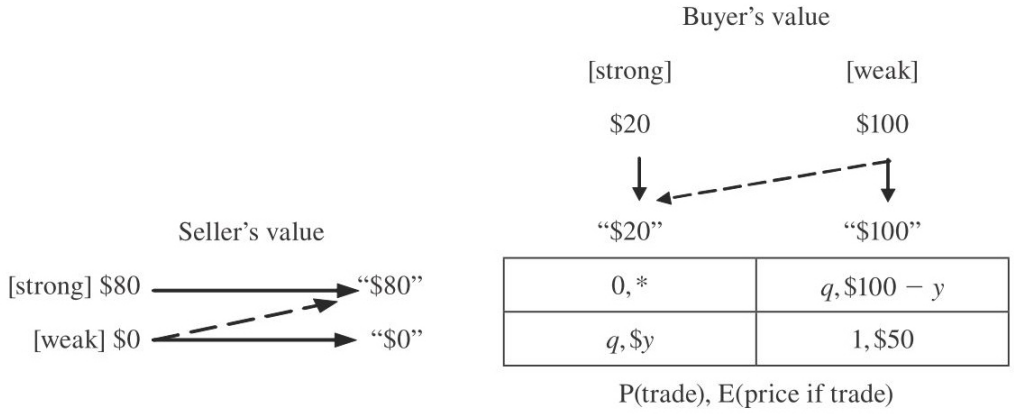

Figure 2.4: Symmetric scheme with parameters q and y. From Myerson (2008).

compatible mechanism must satisfy. For simplicity, we make some assumptions, which are shown in the table in Figure 2.4. As before, it will be assumed that if the seller and the buyer are both strong, the probability of trade is 0, and if both are weak, the item will be sold for \$50 with probability 1. If one is weak and the other strong, then the trade will occur with a probability q that does not depend on who is weak or strong. If the trade occurs, the profit of a weak trader against a strong trader is some number y, which again is the same no matter who is weak or strong, as can be seen in the lower-left and the upper-right cell of the table. The buyer and the seller are thus treated symmetrically. (The assumptions are only for simplicity. See Myerson (2008) p. 595 for how they can be relaxed.)

With these assumptions in place, we can ask what constraints the parameters q and y must satisfy in an incentive-compatible mechanism. First, note that a strong buyer would agree to buy the item only for a price smaller than (or equal to, suppose) \$20. Similarly, a strong seller would sell the item only for a price larger than or equal to \$80. Therefore, the parameter y must satisfy the participation constraint $y \leq 20$.

For honesty to constitute an equilibrium, we must make it an optimal response for any trader to honestly reveal her type if she expects the other trader to

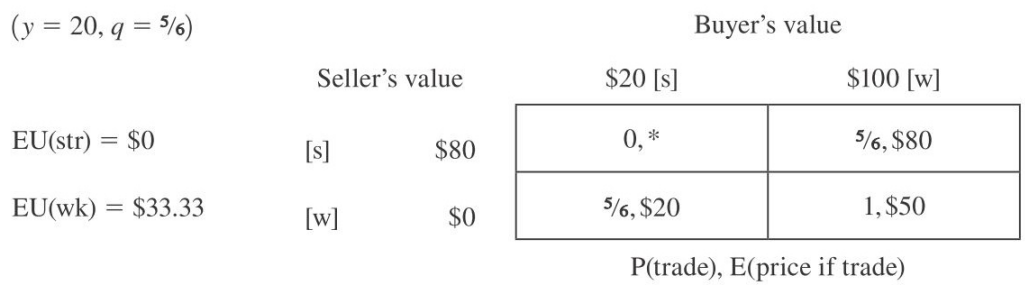

Figure 2.5: The 5/6-mechanism, which is incentive-compatible. From Myerson (2008).

honestly reveal her type too. It can be verified in Figure 2.4 that a strong seller or buyer would never gain by claiming to be weak. But, depending on the parameters y and q, a weak seller or buyer might gain by claiming to be strong. Consider the weak buyer. His expected payoff from honesty is $1/2(q)(y) + 1/2(50)$, and his expected payoff from lying is $1/2(q)(100 - y)$. So for honesty to be an equilibrium, the parameters q and y must satisfy the incentive constraint $1/2(q)(y) + 1/2(50) \geq 1/2(q)(100 - y)$, which reduces to $q \leq 25/(50 - y)$. (It can easily be verified that this constraint is identical for the seller.)

Thus, a mechanism that satisfies the participation constraint $y \leq 20$ and the incentive constraint $q \leq 25/(50-y)$ makes honest participation an equilibrium and is therefore incentive-compatible. Setting $y = 20$ achieves the largest feasible probability for the trade to happen, viz. $q = 5/6$. This mechanism, call it the 5/6-mechanism, is shown in Figure 2.5. The expected profits (ex ante, that is, before the types are revealed) for each trader in this mechanism are $[0 + 0 + (5/6)20 + (1)50]/4 = 16.67$.

This concludes the example. It shows how incentive compatibility delimits the amount of social welfare that is feasible: the 5/6-mechanism yields a positive probability that the item will not be traded when the traders are of different types, even though this means that the item will not go to the individual who values it most. So ex post, after the traders reveal their types,

the mechanism produces allocative inefficiencies in cases in which the traders are of different types but the trade does not occur. We saw that there is no incentive-compatible mechanism with a lower probability of such allocative inefficiency than the 5/6-mechanism.[10] Thus, this mechanism determines the boundary of feasible social welfare when agents are strategic.

The same point can also be made by comparing expected profits. There is no incentive-compatible mechanism that would give both traders a higher expected profit than the $16.67 that the 5/6-mechanism yields: in technical terms, this mechanism is *ex ante incentive efficient*. If it were possible to rely on the players' honesty and dispense with incentive compatibility, then the Split-the-Difference mechanism would yield an expected profit of $17.5 for both players ($[0 + 10 + 10 + 50]/4 = 17.5$), thus improving upon the incentive efficient mechanism. However, institutional designers do not expect rational and intelligent people who will interact through the mechanism to reveal their private information honestly unless it is in their best interest to do so. Therefore, they treat incentive compatibility as a general constraint on the institutions that govern social interactions. Doing so delimits the social welfare that institutions can achieve.[11]

It is important to note that, while we introduced the problem of incentive compatibility with regard to agents revealing their private information, the problem is a very general one facing social planners. Of every institution whose outcomes depend on individuals revealing private information or performing actions that the social planner cannot fully observe, it can be asked whether it gives those individuals incentives to reveal their information and to perform their hidden actions obediently. As an example of an incentive-*in*compatible institution concerning hidden actions, remember the rats in Hanoi: a failure to

[10] In fact, a result from mechanism design, the "revelation principle", implies that, if the traders are strategic, there is no mechanism (incentive-compatible or not) in which the probability of allocative inefficiency is lower than in the 5/6 mechanism.

[11] These limits to social welfare can be examined more generally: see Holmström and Myerson (1983).

provide incentives for the locals to perform the "right" actions, which the colonial rulers could not observe (or which they did not want to observe, to avoid descending to the sewer tunnels), caused the rat population to spread. Since it is hard to think of an institution that does not rely on individuals' private information and in which all the individuals' actions can be fully observed, the problem of incentives is ubiquitous in institutional design.

2.3 Non-Classical Theories Conflict with the Constraint of Incentive Compatibility

Bringing together non-classical theories of rationality and incentive-compatible institutional design, we arrive at a conflict: what institutional designers expect individuals to do is inconsistent with what non-classical rationality requires them to do, at least under certain conditions. In the example from the last section, if the two traders reasoned as a team, or constrained their maximization, they would commit to being honest about their types in the Split-the-Difference mechanism and they could reap the benefits from this commitment, receiving an expected payoff of \$17.5. The constraint of incentive compatibility reduces the traders' expected payoff to \$16.7 at best. So if the institutional designer's goal is to maximise the traders' expected payoff, she would only implement an incentive-compatible mechanism if she does not expect the traders to be honest when they can profit from lying in Split-the-Difference.

In general, when an institutional designer imposes incentive compatibility as a constraint, she does not believe that the individuals interacting through the institutions reason as a team, or constrain their maximization: for if she did, she would expect that more social welfare could be achieved when dropping the constraint. Thus, the following three claims form an inconsistent triad: (i) institutions should be designed so as to maximise social welfare; (ii) they should be designed incentive-compatible; and (iii) the individuals interacting

through the designed institutions are non-classically rational. It is easy to see that any two of the claims imply the negation of the remaining claim. Asserting that institutions should be designed so as to maximise social welfare and requiring incentive compatibility (that is, asserting (i) and (ii)) implies that individuals interacting through these institutions are not non-classically rational (that is, denying (iii)). Likewise, stating that institutional design should maximise social welfare and that individuals are non-classically rational ((i) and (iii)) implies that they should not be designed incentive-compatible (not (ii)). And finally, stating that institutions should be designed incentive-compatible and that individuals are non-classically rational ((ii) and (iii)) implies that institutions should not be designed so as to maximise social welfare (not (i)).

In order to avoid inconsistency, non-classical theories must reject one of (i), (ii) and (iii). But rejecting (i), that institutions should be designed so as to maximise social welfare, should be excluded as an ethically unreasonable position. Thus, these theories are left with two options: they could reject that institutions should be designed incentive-compatible (reject (ii)), or that individuals interacting through the designed institutions are non-classically rational (reject (iii)). There are two ways in which (iii) could be rejected because one could negate either the "rational"-part, or the "non-classical"-part of the claim. That is, a proponent of a non-classical theory of rationality could state that the individuals interacting through the designed institutions are *irrational*, where the meaning of "rational" is fixed by her non-classical theory. Or she could state that these individuals are rational, but not in any non-classical meaning of the word. So in the latter case, proponents of non-classical theories would concede that their own theories do not apply when it comes to individuals interacting through institutions.

In summary, in order to avoid inconsistency and assuming that we should design institutions to maximise social welfare, non-classical theorists must make one of the following claims:

(1.) Incentive compatibility is not required as a constraint on institutional design;

(2.) the individuals interacting in the designed institutions are irrational; or

(3.) non-classical theories of rationality do not apply to institutional design.

I shall next consider these options in turn.

2.3.1 Away with Incentive Compatibility?

Since Leonid Hurwicz coined the term "incentive compatibility" and gave it a rigorous treatment using game theory (cf. Hurwicz (1972)),[12] it has become an almost universally acknowledged constraint on institutional design, and the prospects of relinquishing it appear dim to most. However, some critics of classical rationality more or less explicitly dissent from the view that incentive compatibility is required as a constraint on institutional design. For instance, Amartya Sen criticises

> "...the assumption that when asked a question, the individual gives that answer which will maximize his personal gain. How good is this assumption? I doubt that in general it is very good...What is at issue is not whether people invariably give an honest answer to every question, but whether they always give a gains-maximizing answer, or at any rate, whether they give gains-maximizing answers often enough to make that the appropriate general assumption for economic theory." Sen (1977), p. 331-2

[12] Ancestors of the concept of incentive compatibility date back at least to David Hume, who famously wrote: "In contriving any system of government and fixing several checks and controls of the constitution, every man ought to be supposed a knave and to have no other end, in all his actions, than private interest. By this interest, we must govern him, and by means of it, notwithstanding his insatiable avarice and ambition, cooperate to the public good" (Hume (1742)). Thus, according to Hume, institutional designers must direct self-interested actions to the public good. Moreover, Adam Smith's praise of the market mechanism, which presumably achieves socially desirable outcomes through agents' pursuit of their self-interest, is consistent with incentive compatibility too (Smith (1776)).

Sen is concerned with the provision of public goods in this passage, where everyone should pay a price proportional to how much they value the good in question. The conventional logic of incentive compatibility implies that people must be given incentives to reveal their true valuations, otherwise they would pretend to have low valuations in order to pay less. But Sen believes that people sometimes "commit" to being truthful, which would render incentive compatibilty moot. Similarly, Gauthier not only proposes constrained maximization as a theory of rationality; he also asserts that people do often constrain their maximisation. As we have seen, incentive compatibility sacrifices social welfare compared to some incentive-*in*compatible mechanisms where everyone acts obediently. So in order to maximise social welfare, a social planner *should not* impose incentive compatibility as a general constraint on institutional design, according to this view.[13] [14]

There are at least two problems with this view. The first is empirical evi-

[13] I take the following quote as evidence for this interpretation of Gauthier, although it is not entirely clear whether he has in mind incentive compatibility.

> "We pay a heavy price, if we are indeed creatures who rationally accept no internal constraint on the pursuit of our own utility ... Could we but voluntarily comply with our rationally undertaken agreements, we should save ourselves this price.
>
> We do not suppose that voluntary compliance would eliminate the need for social institutions and practices, and their costs. But it would eliminate the need for some of those institutions whose concern is with enforcement. Authoritative decision-making cannot be eliminated, but our ideal would be a society in which the coercive enforcement of such decisions would be unnecessary. More realistically, we suppose that such enforcement is needed to create and maintain those conditions under which individuals may rationally expect the degree of compliance from their fellows needed to elicit their own voluntary compliance. Internal, moral constraints operate to ensure compliance under conditions of security established by external, political constraints." Gauthier (1987), p. 164-65.

[14] Perhaps Luigino Bruni and Sugden (2013) can also be interpreted in line with Sen and Gauthier. They criticise the standard view of why markets bring about mutual benefits to buyers and sellers under ideal conditions. Following Adam Smith, these benefits are typically interpreted as the unintended consequence of the pursuit of the traders' individual self-interest (Smith (1776)). Against this tradition, Bruni and Sugden argue that virtuous traders *intend* this mutual benefit, instead of their own benefit only. In our example, if the buyer and seller intended mutual benefit, this might motivate them to be honest about their types in Split-the-Difference, thus suggesting that incentive compatibility should not be required.

dence. Whether we retain or let go of incentive compatibility should depend on whether incentive-compatible institutions can achieve more social welfare than some class of incentive-*in*compatible institutions can achieve. In principle, this can be resolved empirically, by comparing the extent to which people's behaviour approaches classical (or non-classical) rationality when they interact through institutions. McFadden (2009) surveys some applications of mechanism design and shows that individual behaviour meets the expectations of classical rationality when incentives are large, even in complex choice settings. In contrast, when incentives are small and ambiguous, the individual deviations from these expectations grow, which McFadden attributes to individuals' putting less effort into determining best replies, and being more distracted by irrelevant factors. These findings suggest that incentive compatibility is crucial, especially when much is at stake for the individuals who interact through the institution in question, whereas there might be some room for relaxing this constraint in some smaller-stake contexts.

It is unlikely that McFadden's study alone could convince proponents of non-classical theories who believe that incentive compatibility is not required as a constraint on institutions. But there is a second problem for their view, which concerns the burden of proof. As we have seen, incentive compatibility delimits feasible social welfare. But incentive-*in*compatible institutions produce adverse incentives that may diminish welfare much more. Remember the rats in Hanoi.[15] Arguably, because of the potentially large welfare losses that are at stake, a social planner should require incentive compatibility as a default constraint on institutional design unless there is very good evidence that we can let go of it in particular institutional arrangements. So the burden of proof lies on the likes of Gauthier and Sen who believe that we should let go of it.

[15] This point can be made more precise in the example from the last section. We saw that the incentive-compatible, 5/6-mechanism yields both players an expected profit of \$16.67. This is the maximum feasible social welfare when the traders are strategic. In contrast, when players are strategic in the incentive-*in*compatible, Split-the-Difference mechanism, there is a symmetric equilibrium in which both traders falsely report strong types when they are weak with a probability of 3/5. The expected payoff in this equilibrium is only \$10 for both traders.

The latter point reinforces the weight of the empirical study above. The study provides evidence, at the very least, that proponents of non-classical theories will have a hard time to live up to their burden of proof. In combination, these two arguments go some way towards urging that option (1.), that is, doing away with incentive compatibility, should be resisted. It seems to be a case of utopian thinking that we could let go of incentive compatibility, with potentially catastrophic consequences in high-stake situations. In Ken Binmore's words, "[i]n rejecting the second-best outcome in favor of an illusory first-best outcome, you condemn yourself to a third-best or worse outcome" (Binmore (2007), p. 31).

2.3.2 Institutions for the Irrational?

A non-classical theorist could reason as follows: "Individuals interacting through institutions should not be expected to team-reason, or to constrain their maximization. Instead, these individuals can be expected to follow best-reply reasoning, which justifies retaining incentive compatibility as a constraint on institutions. Because the individuals cannot reap the fruits of cooperation, they are *irrational*." This is option (2.) for harmonising non-classical rationality with the constraint of incentive compatibility.

However, assuming people's irrationality is unpopular, and our imagined non-classical theorist is likely to be a fiction. There is a broad consensus that we ought to treat people as rational and that the design of our institutions and other policies should reflect this. A failure to do so is seen as a lack of respect towards fellow citizens. For instance, nudging, which seeks to manipulate people's decisions by altering the presentation of their available options, is sometimes criticised because it allegedly fails to treat people as rational, deliberative agents (for a discussion, see Bovens (2009), p. 209 et seq.). What is more, many proponents of non-classical theories explicitly defend the view that human beings can, and often do, meet their rationality standards. For

instance, Gauthier (1987) endorses the "conception of human beings as rational (or potentially rational) individual actors" (p. 93). In fact, qualms with the classical assumption that rational individuals cannot achieve optimal outcomes in certain, e.g. PD-like situations, are the very reason for proposing non-classical theories in the first place. Non-classical theorists demand that stringent rationality requirements be applied, and it would be an odd view to demand this while holding that people cannot in fact meet these requirements.

Thus far, the option of assuming people's irrationality looks rather dismal. But perhaps this is premature. I have until now talked as if non-classical theorists presume that agents are irrational whenever they do not constrain their maximization, or do not reason as a team. But at least for some versions of non-classical theories, the assumption would suffice that only some, not all, agents are irrational, in order to justify the constraint of incentive compatibility. For example, remember that a constrained maximizer will rationally defect in PD games when a sufficiently large fraction of the agents with whom she interacts are defectors, and depending on the level of translucency. Similarly, if a fraction of agents renege on their commitment to reveal information honestly or to act obediently in an incentive-*in*compatible institution, it might be rational for the others to do the same, according to constrained maximization. For some varieties of team reasoning, similar arguments could be constructed. For instance, according to Sugden (2015), in order to team reason, a player requires assurance that the other player(s) will play as team members as well – which is perhaps not the case if other players are irrational. Anticipating that non-classically rational people would defect in an incentive-*in*compatible institution in the presence of non-classically irrational people, proponents of these theories could argue that the institution better be designed incentive-compatible.

This line of argument nevertheless commits proponents of non-classical theories to the assumption that *some* people are irrational, which may be deemed

undesirable. Alternatively, non-classical theorists could retain incentive compatibility for a similar reason for which drivers wear seatbelts: we don't expect the accident to happen, but why take a chance? Similarly, imposing incentive compatibility would be the consequence of a kind of precautionary reasoning: "There *might* be a fraction of irrational people who won't cooperate. In the face of this uncertainty, better take preventive action and impose incentive compatibility as a constraint on institutional design, because failing to do so entails the risk of bringing about a socially undesirable outcome."

Perhaps some proponents of non-classical theories will consider the latter strategy for justifying incentive compatibility to be a desirable option. Others may reject it on the basis that assuming people are *potentially* irrational still involves an undue paternalistic attitude towards them. We need not take up a stance on this matter. We simply note that there is an asymmetry between classical theories of rationality and non-classical theories when combined with this strategy: while for the former, incentive compatibility is imposed *because* of the assumption that people are rational, for the latter, incentive compatibility is imposed because people are potentially irrational and *despite* the expectation that they are rational. For anyone committed to the assumption that people should be treated as rational, this constitutes an important disadvantage of non-classical theories when combined with this strategy relative to classical theories.

2.3.3 Non-Classical Theories do not Apply to Institutional Design

If a proponent of a non-classical theory is reluctant to give up incentive compatibility as a constraint on institutional design and to presume that institutions are designed for (potentially) irrational people, this seems to be a self-defeating position, at least at first glance: he wants incentive compatibility and he wants to treat people as rational, but his rationality standards

don't allow for both. His remaining option is therefore to concede that rational individuals do not constrain their maximization and do not reason as a team when they interact through institutions. In other words, he would concur that his own rationality standards do not apply to institutional design. Plausibly, he would instead assert that the standards of classical rationality apply. He would seem to overthrow his rationality standards and become a proponent of a classical theory.

A more positive interpretation would be to allow the possibility of a contextual account of rationality. One could endorse a classical notion of rationality when it comes to agents interacting through institutions. One could at the same time welcome the possibility that there might be conditions under which non-classical theories do apply. But this strategy immediately raises a further question: what are these conditions, and in which contexts are they satisfied? Without an answer to this question, a contextual theory of rationality remains indeterminate.

Some theorists have provided conditions under which individuals are supposedly prone to act in accordance with the predictions of non-classical rationality. Plausibly, these theorists would also maintain that, when an interaction satisfies these conditions, then this interaction should be governed by non-classical rationality. The conditions they provide appeal either to structural features of games, or to features outside the game-theoretical description of an interaction. For instance, consider Bacharach's account, which includes both kinds of conditions. Concerning the structural features, he suggested that "strongly interdependent" games increase the likelihood of individuals identifying as a team. Strongly interdependent games are games with strategy profiles (which are not necessarily equilibria) that strictly Pareto-dominate some pure equilibria (Bacharach (2006), p. 84 et seqq.). Concerning external features, Bacharach took findings from social psychology to show that belonging to the same social group, face-to-face contact, and other factors can trigger

team reasoning (ibid., p. 76 et seqq.).

However, Bacharach's conditions fail to pick out only those contexts that could possibly be governed by non-classical rationality because they do not exclude institutional contexts, which are governed by classical rationality, according to the contextual account of rationality that we are interested in here. Concerning his structural feature, many games that are induced by incentive-*in*compatible mechanisms have strategy profiles that dominate some equilibria and thus they constitute strongly interdependent games. Split-the-Difference is an example of such a mechanism. If some of these mechanisms are institutions, the criterion of strong interdependence does not serve to pick out the conditions under which individuals should be expected to team-reason.[16] Below, I will argue more generally that structural features of games are inapt for excluding institutional contexts, because structural features do not go a long way towards defining institutions.

If this analysis is correct, proponents of a contextual account of rationality are left with external features. As with structural features, many of the proposed external features – including belonging to the same social group or having face-to-face contact – could well be satisfied when individuals interact through institutions. In general, any attempt to provide conditions for the applicability of non-classical rationality will face this difficulty: the problem is not only to find conditions under which non-classical theories should presumably apply, but which are not met whenever agents interact through institutions. A possible response to this problem is to exclude institutions by fiat from the proposed conditions. For instance, instead of claiming that non-classical rationality should apply in face-to-face interactions, it could be claimed that it

[16] Some accounts of institutions exclude incentive-*in*compatible mechanisms as possible candidates for institutions. I will argue in the next section that this is a mistake, but even if incentive-*in*compatible mechanisms could not be institutions, Bacharach's criterion won't do. The reason is that incentive compatibility is required as a constraint because individuals should not be expected to reason as a team in incentive-*in*compatible mechanisms, strongly interdependent or not. But if strong interdependence were to trigger team reasoning, incentive compatibility would not generally be required, because strongly interdependent, incentive-*in*compatible mechanisms could be trusted to maximise welfare.

should apply in such interactions in contexts other than institutions. However, this would merely shift the question one step back, as it would require an answer to the question what are the special features of institutions that delimit the possible scope of non-classical theories in this way.

Furthermore, using external features in order to single out those contexts in which non-classical theories of rationality in games should presumably apply is problematic in its own right. It is typically difficult to confirm that these features could trigger team reasoning, or other non-classical reasoning. For instance, if cooperative outcomes become more likely in inter-social-group interactions, this might be because players sympathise more with each other, not because they reason as a team. But sympathy is a motivationally relevant factor that affects individuals' utilities, and is as such orthogonal to non-classical rationality.

In summary, if proponents of non-classical theories choose option (3.), viz. concede that their rationality standards do not apply to people's interactions through institutions, prima facie this looks like bad news for their theories. In response, they can advance a contextual notion of rationality and insist that non-classical rationality standards apply in some contexts in which individuals do not interact through institutions. But they are then asked to come up with conditions under which non-classical theories presumably apply. Some proposals have been made that might be interpreted as providing such conditions, but these proposals fail to pick out conditions that could not be satisfied when agents interact through institutions. Thus, it appears fair to conclude that no convincing account has been proposed to render option (3.). Moreover, providing conditions that could not be satisfied when agents interact through institutions seems to pose a general difficulty.

This is not a principled argument against the viability of this option; it merely shows that it is difficult to see how a convincing account could be given. Because this option implies that institutions delimit the possible scope of non-

classical theories, I shall next make a few clarifying remarks on the concept of institution, which we have treated as primitive until now.

2.4 Some Remarks on the Concept of Institution

"Institution", in its everyday usage, is a somewhat enigmatic term that can refer to such diverse things as a society's morality, economic systems, free trade agreements, the Roman Catholic Church, or Liverpool FC. I do not intend to give an account that uncovers the fundamental structure that all institutions have in common – if there is such a thing. Nevertheless, my remarks may have implications for attempts to provide accounts of institutions, which are supposed to do just that. An institution, as I have used the term here, is an entity that governs some specified interactions through the use of an institutional rule, or *mechanism*. A mechanism specifies how the outcomes of an interaction depend on possible combinations of individuals' actions. For instance, remember the Split-the-Difference mechanism, depicted in Figure 2.3. It specifies how the probabilities and prices at which the trade happens (the outcomes) depend on the information that the two traders reveal (their actions). Imagine a market that facilitates one-shot interactions between single buyers and sellers of single items through Split-the-Difference. If this market existed, it could be seen as a simple institution.

It is important to note that there is typically more to an institution than a mechanism. For instance, our imagined institution may have a material basis, such as a marketplace where the traders state their valuations, and enforcement devices to ensure that the trade effectively happens through Split-the-Difference, and not some other mechanism (e.g. "Wild-West", in which the strongest "trader" simply takes the item). However, the theoretical literature on institutions often neglects their material bases,[17] and I will temporarily do

[17] Cf. Rabinowicz (2018), who criticises this neglect.

this too, in order to connect to this literature.

The conception of institutions as mechanisms might be seen as akin to a tradition that interprets institutions as *rules*. But rules, in the latter sense, are interpreted simply as players' actions or strategies, because they can be formulated as prescriptions of the form, "choose X", or "do Y" (see Hindriks and Guala (2015a), p. 463, or Guala (2016), p. xxiv et seq.). Under this interpretation, a rule is not a mechanism, but a subset of a mechanism (which, to repeat, specifies all the actions available to the players, as well as what outcomes the possible combinations of actions result in). Guala and Hindriks criticise the rules-account because it fails to explain why successful institutions manage to influence behaviour in desirable and predictable ways. For instance, some rules are de facto ineffective, not followed by anyone: "[t]raffic lights in Milan are regulation, in Rome they are a suggestion, and in Naples they are just decoration" (Guala (2018a), p. 541). This is the case even though the formal traffic rules (e.g. "stop at a red light") are identical in the three cities. Guala and Hindriks emphasise that effective rules are those that agents have incentives to follow. Therefore, they combine the institutions-as-rules account with a second tradition, which interprets institutions as *equilibria*. In the Guala-Hindriks account, institutions are *rules-in-equilibrium*: they prescribe the individuals' actions, just like in the rules account, but at the same time they require that these actions constitute an equilibrium. Since in equilibrium, no one would be better off by unilaterally deviating from playing her part, agents are motivated to follow these rules, which explains why the institution "works".[18]

[18] This is a simplified statement of Guala's and Hindriks's account; see Hindriks and Guala (2015a) and Guala (2016) for more details. In particular, they maintain that institutions solve coordination games with multiple equilibria by providing correlation devices: they view institutions as *correlated equilibria*, instead of Nash equilibria. For present purposes, the coordination aspect that many institutions undoubtedly involve can be ignored. As to correlated equilibria, it has been argued that Nash equilibria are more suitable to account for institutions (cf. Binmore (2015) and Rabinowicz (2018)). In response, Hindriks and Guala (2015b) agree that "the unified theory could be expressed in terms of Nash equilibria" (p. 516). So correlated equilibrium appears to play a less important role in their theory than they initially asserted.

I would like to raise two issues about this account. The first is a suggestion, namely that mechanisms should form an essential part of an institution in the rules-in-equilibrium account. In an equilibrium, players choose optimally given their (correct) expectations about the other players' behaviour. But expectations, as well as optimal responses to them, essentially depend on the mechanism in effect. This is why social planners design mechanisms, not only rules: they do not simply tell individuals what to do, but they seek to determine the consequences of their actions, including consequences of deviations from the intended actions. Reference to mechanisms, instead of to rules alone, would therefore increase the explanatory power of the rules-in-equilibrium account, because mechanisms provide information about *why* a rule is, or is not, in equilibrium. For example, consider again our example of the sale of the item. The relevant rule in this example is, "be honest about your type". According to the rules-in-equilibrium account, this rule could only form an institution if it is in equilibrium. But whether it is depends on the mechanism used: the 5/6-mechanism implements this rule in equilibrium while Split-the-Difference does not. Thus, explaining why a rule could constitute an institution in the sense of the rules-in-equilibrium account requires reference to the mechanism, which this account omits. Or consider that, according to the rules-in-equilibrium account, traffic lights are an institution in Milan, but not in Naples. This is because different equilibria are in effect in the two cities, of which only one coincides with the relevant rule, "stop at red". That different equilibria prevail can be due to either of two things: either the identical mechanism is effective in the two places but a different equilibrium selection took place, or the difference is due to different mechanisms effective in the two cities, despite the identity of the rule (for instance, disobeying the rule might lead to different consequences in Milan, e.g. a higher probability of a fine, than it does in Naples). In either case, an explanation of the difference that simply maintains that different equilibria are in effect remains sketchy; a complete explanation requires reference to a mechanism.

The second issue concerns the argument for the rules-in-equilibrium account that stating a rule is not sufficient for an institution because a rule that does not make the actions that it prescribes an equilibrium is ineffective. This argument only repeats that, in order to entice individuals to perform certain actions, institutions must be incentive-compatible, that is, these actions must form an equilibrium of the mechanism that governs the institution. Thus, institutions in the Guala-Hindriks sense can be thought of as incentive-compatible institutions in the sense of mechanism design (approximately that is, for see the preceding footnote). Which definition is preferable? This is mainly a matter of convention, but I would like to point out two apparent problems for the rules-in-equilibrium account. Both stem from the fact that this account treats institutions as incentive-compatible by default. First, the rules-in-equilibrium account cannot accommodate a kind of institutional failure, namely, that many institutions fail *because* they produce adverse incentives. This implies that they were institutions in the first place. But according to the Guala-Hindriks account, incentive-*in*compatible institutions are not badly designed institution, which may fail for this reason, but they are not institutions at all.[19] This sits uneasily with the standard use of the word "institution", which allows for the possibility of this kind of institutional failure. For instance, the failure of socialist economies has been attributed to the incompatible incentives that those economies provide their citizens with (Myerson (2009)). But surely, economic systems are prime examples of institutions in the standard meaning of the term. Denying them the status of institution seems problematic, unless there is good reason to do so.

It would be of no help to reply that socialist economies can be excluded as

[19] At some points, Guala could be interpreted differently. For example, he writes that "an effective institution is an equilibrium state where all the relevant individuals have an incentive not to deviate from a certain pattern" (Guala (2016), p. 18; also cf. Guala (2018a), p. 541). This passage could be interpreted as saying that, when not all relevant individuals have an incentive not to deviate from the pattern in question, we don't have an effective institution (as opposed to no institution at all). However, this interpretation is inconsistent with his definition that institutions (simpliciter) are rules-in-equilibrium, which implies that out-of-equilibrium states are not institutions.

important institutional arrangements as most socialist economies have vanished. This is because, if Myerson's analysis is correct, capitalism is incentive-*in*compatible too.[20] This points at the second possible problem for the rules-in-equilibrium account: the account risks disregarding the fact that incentive-compatible institutions are in most interesting problems difficult to come by. The simple, two-trader example might convey a taste of this difficulty. Relaxing some of its unrealistic assumptions – e.g. that each trader is of one of two types with an independent probability that is commonly known – quickly makes the problem intractable. The search for incentive-compatible institutions is a research programme that has led to important, general characterisations of the social goals that can be implemented, and what form the mechanisms that implement them take (Maskin (1999)). But an account of institutions that assumes their incentive compatibility by default runs the risk of trivialising the search for incentive-compatible institutions.

However, this does not mean that defining institutions as mechanisms (incentive-compatible or not) would fare better. Instead of excluding important institutions, this would be far too permissive. Not all mechanisms are institutions because any interaction is subject to a mechanism, but not every interaction is an institution. So what are institutions, after all? I have suggested that they are entities that govern specified interactions through the use of mechanisms, and mechanisms can be used to make some important characterisations. In particular, institutions that use incentive-compatible mechanisms are likely to be successful in implementing the social goal in question, while those that use incentive-*in*compatible mechanisms are unlikely to implement the social goal and are prone to fail. But, as mentioned at the beginning of this discussion, there is more to an institution than its mechanism, for instance, its material basis. Not much more of interest can be said about the concept

[20] Myerson (2009) gives simple models to show that the choice between capitalism and socialism is subject to trade-offs between the two kinds of incentive problems: while socialism is better than capitalism in incentivising citizens to reveal their private information, capitalism has an advantage when it comes to motivating hidden actions.

of institution at this level of abstraction.

This discussion has important implications for non-classical theories of rationality. Remember that some non-classical theorists may wish to hold a contextual notion of rationality (that is, they advocate option (3.) of the previous section). Such a non-classical theorist will need to delimit the scope of her theory by excluding institutional contexts. Remember also that she could make use of two kinds of features to delineate the possible scope of her theory: structural features of games, or features that lie outside the game-theoretical description of an interaction. The two kinds of features recurred in our discussion about the concept of institution: some attempts have been made to define institutions through structural features alone, using concepts from game theory, but institutions also possess features, such as a material basis, which are external to their game-theoretical description. If my analysis is correct, structural features do not go a long way towards defining institutions. In order to exclude institutional contexts, it seems that our non-classical theorist must then make use of external features. But, as I argued above, external features are problematic in their own right.

2.5 Conclusion

The colonial rulers' mistake in the "Great Hanoi Rat Massacre" was to ignore the fact that their bounty system did not provide the locals with incentives to combat the rats. We know the end of the story. The failure of the system is an illustration of the importance of incentive compatibility as a constraint on the institutions through which we interact.

Non-classical theories of rationality in games are at odds with incentive compatibility. Therefore, proponents of these theories face three options. They could let go of incentive compatibility, but they would thereby make the same mistake as the colonial rulers. Or they could blame citizens for their purported

irrationality, but this option looks rather unappealing from the perspective of non-classical theories. Or they could decide not to blame the citizens but their own rationality standards. They may argue that these standards apply under conditions that are not satisfied when agents interact through institutions. The latter strategy would then direct attention to the very concept of institution, because institutions would delimit the possible scope of non-classical theories. Since game theory does not go a long way towards delineating the concept of institution, theorists opting for this option need to search for conditions under which non-classical rationality allegedly applies outside the game-theoretical description of a choice situation. It is difficult to see how this could be done.

Institutional design constitutes a challenge for non-classical theories of rationality to which proponents of those theories haven't given a convincing response. In the meantime, the institutions through which we interact must be designed. In doing so, the working hypothesis should be that individuals are rational, in the classical meaning of the word.

Chapter 3

A General Epistemology of Economic Engineering

Economists are increasingly engaged as "engineers": instead of taking market outcomes as mere phenomena to be explained or predicted, they also seek to bring about desirable outcomes by creating or changing markets. While the recognition of this field in economics is growing – for instance with the award of recent Nobel Memorial Prizes[1] – its precise epistemology is less clear. Most philosophers of economics have examined the use of models and other tools in the service of explanations or predictions, and not in how interventions change market outcomes.

The few philosophers of science who have discussed market design – most notably, Anna Alexandrova and Francesco Guala – have noted a complex dependence of this practice on different sources of evidence, in particular experimental methods and game-theoretic models. However, they have only examined a case from spectrum auction design. In this chapter, I introduce a

[1] For example, in 2007, the award went to Leonid Hurwicz, Eric Maskin und Roger Myerson "for having laid the foundations of mechanism design theory" (The Royal Swedish Academy of Sciences (2007)), and in 2012, to Alvin Roth and Lloyd Shapley "for the theory of stable allocations and the practice of market design" (The Royal Swedish Academy of Sciences (2012)).

novel case study: the matching market that allows medical graduates in the US to find training positions in public hospitals. In the 1990s, game theorists were significantly involved in the redesign of this market, which was commissioned as a response to market failure. The result was one of the great success stories of market design, and a key motivation behind the 2012 Nobel Prize to Lloyd Shapley and Alvin Roth, which is worth discussing in some detail.

My case study corrects some of the claims that have been made about the epistemology of economic engineering. In particular, Alexandrova is critical of the role of analytical models, because in the design of spectrum auctions, models were not decisive for some of the most important design questions. Instead, experiments were crucial for informing the final design. She contrasts descriptions of the design process by theoretical and experimental economists, and argues that theorists oversell the contribution of their models. But this picture of theorists and experimentalists as competing for the importance of their contributions does not accurately represent other instances of economic engineering. In the case of the matching market for medical graduates, the theoretical and experimental economists involved do not appear to be competitors; in fact, they were largely the same persons.

I put forward an account of the process by which engineering knowledge is generated, which is consistent with both cases from auction design and matching. According to this account, models, e.g. from game theory, allow defining properties that may correspond to policy goals. Moreover, the theory encodes counterfactual information in its models, which can be used to track how interventions within a model change its outcomes. I use directed graphs, which may acquire a causal interpretation, to illustrate how these model-interventions suggest interventions in the target institutions to implement the policy goals. But models famously make false assumptions and isolate mechanisms which are in the real world interfered with by other, distorting mechanisms, thus precluding a causal interpretation. The complementary application of laboratory,

natural and computational experiments, but also tinkering with the models, allows for inferences about real-world institutions.

Thus, models and empirical methods are complementary tools in the search for desirable institutions. Typically we cannot say what combination of these tools – models, experiments, etc. – is efficient for economic engineering. This urges some scepticism with respect to the so-called "efficiency question in economics": a recent call in the philosophy of science for evaluating whether models are efficient means to achieve economists' goals, relative to other tools. My discussion shows that in economic engineering, this may be the wrong question to ask.

This chapter is organised as follows. In the next section, I present some of the conclusions that philosophers of science have drawn from the design of spectrum auctions about how knowledge is generated in market design. Section 3.2 describes the redesign of the matching market for medical graduates. In section 3.3, it is argued that this case proves inconsistent with parts of the claims that are motivated by the spectrum auctions, and calls for a general account of the epistemology of economic engineering. I endeavor to provide such an account in section 3.4, in which directed graphs are used to describe how models, complemented by experiments and other tools, allow projecting institutions. Section 3.5 concludes with some remarks on the efficiency question.

3.1 Philosophers of Science on Spectrum Auctions

In the most general terms, economic engineers seek to bring about desirable outcomes through suitable institutional design. I shall focus here on market design, in which the institutions are markets and the outcomes are market outcomes. What properties should be considered to be desirable in a given case

involves ethical questions, which will be disregarded here.[2] I shall assume that exogenous policy goals determine what properties outcomes should possess. Typically, market designers seek to exploit agents' conscious, strategic pursuit of their individual goals in order to implement those properties, which explains why models from (non-cooperative) game theory, or rational choice models more generally, are usually thought to play an important role.

Philosophers of science have to date primarily studied the design of auctions for allocating spectrum licenses to telecommunication service providers in the US. Spectrum refers to a range of electromagnetic frequencies, which are used to transmit video, sound and data. In the US, the Federal Communications Commission (FCC) organises the allocation of the licenses. In the 1990s, the FCC decided to replace an inefficient lottery system with auctions. The auctions aimed to achieve specific policy goals, in particular efficiency, that is, to allocate the licenses to the providers that value them most. What kind of auctions would best promote these goals was a complex design question, which was subject to much controversy among the stakeholders before the first auction was conducted in 1994, and the FCC as well as potential bidders consulted economists about crucial design decisions.[3]

The resulting auctions are generally seen as a big success, efficiently allocating the licenses and raising many billions of dollars of revenue for American taxpayers. Moreover, it was supposedly economic theorists, in particular game theorists, who designed them, so they were presented in media and, little surprisingly, by the theorists themselves, as a success story of game theory. For example, R. Preston McAfee and John McMillan, two theorist-consultants,[4]

[2] See Li (2017a) for a general treatment of ethics and market design. We shall discuss two cases of market design that are particularly ethically loaded in depth in the second part of this thesis: a matching market for asylum, and kidney exchange.

[3] For detailed accounts of the history of the market, see the references to Alexandrova and Guala below.

[4] By the time, both were professors of economics, McAfee at the University of Texas, Austin, and McMillan at the University of California, San Diego. McAfee was involved in the design of the auctions working for the wireless telephone service provider AirTouch Communications, and McMillan for the FCC.

wrote: "*Fortune* said it was the 'most dramatic example of game theory's new power...It was a triumph, not only for the FCC and the taxpayers, but also for game theory (and game theorists)' " (McAfee and McMillan (1996), p. 159; in the quote, they refer to *Fortune magazine*, February 6, 1995, p. 3).

Philosophers of science have challenged this received view. Starting with Guala (2001), they showed that it is not game theory alone that should be credited with the successful design, and they have instead emphasised the importance of laboratory experiments (also cf. Alexandrova (2006, 2008); Alexandrova and Northcott (2009); Guala (2005, 2006, 2007)). They typically note that no theorem from auction theory (a subfield of game theory) was directly applicable to the design of the auctions. The main problem was that bidders value licenses more depending on whether they also get complementary licenses. For example, this could be licenses in a neighbouring state for a bidder who wishes to extend coverage. But different bidders may prefer different bundles of licenses, which is why the FCC could not simply define all the bundles of complementary licenses. Most of the bundles had to be determined through the participants' bidding instead. However, there were no analytical solutions as to what kind of auctions could achieve efficient allocations of goods that include complementarities.

Complementarities were but one complication that analytical models alone could not resolve. Another one was the 'winner's curse': a phenomenon to be prevented, in which the bidder who most overestimates the value of a good wins the auction and thereby makes a loss. There was some evidence, both theoretical and experimental, that open instead of sealed-bid auctions could help reduce the risk of the winner's curse (because bidders can learn about the true valuation by observing other bidders). However, whether this would turn out to be true in the presence of complementarities, and further complications of the market, was not clear. The problem was to find out whether and how these features would interact – and again, models from auction theory were

silent on the matter.

It required experimentalists to cope with these complications. Most prominently, Caltech economist Charles Plott was involved in the auctions consulting for the telecom provider Pacific Bell. Together with his team, Plott created *experimental testbeds*, controlled laboratory environments of auctions in which some features, such as complementarities, can be controlled for. Importantly, testbeds are not intended to isolate single causal mechanisms – unlike models, according to prominent accounts such as Nancy Cartwright's, (e.g. Cartwright (2009)), or Uskali Mäki's (e.g. Mäki (2009, 2011)). Rather, they test material environments holistically, in which possible interactions between different causal mechanisms are in effect. It was numerous such environments in the lab that were finally decisive in favouring a simultaneous, multiple round ascending bid auction over its rivals.

Alexandrova (2006) and Alexandrova and Northcott (2009) contrast the lines in which theoretical economists (such as McMillan) and experimental economist (Plott) interpreted the design process. According to them, the theorists overstate their case when they claim that the FCC "chose an innovative form of auction ... because theorists predicted it would induce more competitive bidding and a better match of licenses to firms" (McAfee and McMillan (1996), p. 160). Instead, they argue that game theory merely provided heuristics and pointed to problems that could possibly arise and which experimentalists should take into account. According to them, the bulk of the evidence required was delivered by experiments. Thus, instead of providing a success story of game theory, the case in fact shows how limited the theory is.

This conclusion can be interpreted in light of a recent proposal by Robert Northcott. Northcott (2018) urges philosophers of economics to engage in what he labels "efficiency analysis": do models, or other tools, provide efficient ways to achieve goals such as explanation, prediction, or in our case, design? Should economists engage more in modelling, or should they invest

resources (money, intellectual labour, education of students, journal space, etc.) elsewhere? The analyses by Alexandrova and Northcott would suggest a revisionary move: since theorists overstate their case, and experimental methods were more important than the narrow models from auction theory, resources should be redirected from theoretical to empirical work in the context of economic engineering.[5] I will next introduce in some depth a case study that hasn't been considered in the philosophy of science, which indicates some caution is needed with respect to this view.

3.2 The Matching Market for Medical Residents

Before becoming doctors, medical graduates in the US are required to take up training positions in public hospitals. The positions are called "residencies", which allow the "residents" to specialise in a specific medical branch. The National Resident Matching Program (NRMP) is the clearinghouse that organises the matchings between residents and residencies. Since the residencies determine a good deal of the residents' future careers, and for the hospitals they provide a significant source of labour force, it is vital that this market is organised fairly and efficiently. In the 1990s, there was a crisis of confidence in the market of the prospective doctors, and the NRMP directors commissioned game theorist and later Nobel laureate Alvin Roth to direct a redesign of this market. I shall sketch the history of the market first, which is based on Roth's and his collaborators' accounts.[6]

In general terms, the matching process is the following. After interviews take place, the NRMP collects rank order lists ("ROLs") from both students and hospitals: lists that reflect the students' preferences over the hospitals they

[5] Alexandrova and Northcott (2015) make a similar, revisionary recommendation in a different context, which is repeated in Northcott (2018). I therefore believe that this is a fair interpretation of their convictions.

[6] Roth's interest in the market dates back at least to Roth (1982). For technical and historical accounts, cf. Roth (1984); Roth and Sotomayor (1990); Roth and Peranson (1999); Roth (2002, 2003); Kojima et al. (2013); Roth (2013, 2015b, 2018), amongst others.

had previously had an interview with, and the hospitals' preferences over the students they had interviewed. The assignments are then determined algorithmically through a *matching mechanism* (see below).

In its early years in the 1950s, the market functioned to the satisfaction of the participants, as indicated by high rates of participation in the system (over 95%; participants are free to decide whether to find matches through the centralised clearinghouse or on their own). However, over the years several changes occurred. Initially, interns were predominantly male, and when female interns entered the market in the 1970s, there were increasing numbers of married couples who graduated from medical school together. Members of couples often have interrelated preferences, particularly to find positions close to one another. For example, even if a member of a couple prefers, say, a position in Boston to a position in Los Angeles other things being equal, these preferences may switch if their partner attains a position close to L.A. In other words, couples give rise to complementarities, similar to those encountered in the FCC auctions. But the matching mechanism in use could not accommodate such desires because it would process only single preference lists. The NRMP modified the system to permit couples to hand in pairs of ROLs together and to specify a "leading member". The mechanism would then match the leading member first, followed by an editing of the other member's preference list to eliminate positions far from that of the leading member. However, this rather ad-hoc modification could not prevent rates of participation from dropping.

The accommodation of couples was not the only challenge the NRMP was facing. Hospitals may have interlinked numbers of positions such as, say, five in the neurology department if internal medicine fills all its positions, but fewer otherwise. This was a source of a different kind of complementarities (cf. Roth and Peranson (1999) for a full description of the kinds of complementarities present in this market). Furthermore, the numbers of graduates relative to

residencies offered increased substantially over the years, which led to matchings being less favourable for the former. In the 1990s, the dissatisfaction among applicants – as expressed by various student associations – was at a peak. Many claimed that the mechanism would show favouritism to the hospitals at their expense; and there was a rumor among applicants that one could 'game the system' by submitting ROLs that wouldn't truthfully reflect their preferences. As a consequence, some student associations requested a change of the matching mechanism, or that the applicants be given more information on how to hand in their ROLs strategically.

The Board of Directors of the NRMP reacted in 1995 and commissioned the design of a new mechanism. They set three policy goals, which the new mechanism should implement as far as possible: to incentivise the applicants and hospitals to stick to the matchings (i.e. not to make arrangements outside the system); to make the matchings as favourable as possible for the applicants; and to reduce their opportunities for strategic behaviour. The new mechanism, which is now known as the "Roth-Peranson algorithm"[7], was first introduced in 1998. It has been working successfully since, and has been adopted by numerous labour market clearinghouses.

In order to answer how Roth and his collaborators reformed the market, we need to dive a bit deeper into the models and other tools used. I shall next sketch a simple model of the market – a model from a subdiscipline of game theory called "matching theory" – and some of the theoretical results that hold in this model. Subsequently, I will flesh out three lessons about how this model was manipulated, enriched, and complemented with other tools to inform the reform of the market.

[7]Elliott Peranson is founder and president of the National Matching Services Inc., a company devoted to providing matching solutions by implementing what they advertise as a "Nobel Prize acclaimed algorithm".

3.2.1 A Simple Model of the Market

From a game theory perspective, a matching mechanism, together with the agents' (that is, the applicants' and the hospitals') preferences, defines a game, in which their actions are to submit ROLs (or to opt out). More formally, there is a set of students $S = \{s_1, \ldots, s_m\}$ and a set of hospitals $H = \{h_1, \ldots, h_n\}$. Each hospital h_i offers a number of residencies which is specified by a quota, q_i. We assume that the agents' preferences $\{\succ_{s_1}, \ldots, \succ_{s_m}, \succ_{h_1}, \ldots, \succ_{h_n}\}$ are transitive, irreflexive and complete lists for each student over the hospitals she had an interview with and that she finds acceptable, and for each hospital over the students it had interviewed and whom it finds acceptable.[8] The agents' actions – their ROLs – are structures just like their preferences: transitive, irreflexive, complete lists over acceptable partners on the other side of the market. Note, however, that agents can be strategic, viz. submit ROLs that do not truthfully reflect their preferences.

A matching mechanism is a function from combinations of ROLs to matchings, which are the outcomes of the game. Formally, a matching μ is a subset of $S \times H$ such that any student appears in at most one pair (i.e., is either matched or unmatched) and each hospital h_i appears in at most q_i pairs (i.e., is either full or has empty places). Let's have a look at the mechanism in use by the time the NRMP directors commissioned the new design. As shown in Roth (1984), in our simple model it is equivalent to the *hospital-proposing deferred acceptance algorithm* (DAA^H). It therefore suffices to sketch the latter:

- In the first step, each hospital "proposes"[9] to the highest-ranked students

[8] This simple model neglects the possibility that some groups of residents may be complementary for hospitals (e.g., a hospital may prefer applicant s_1 to s_2 if it also employs s_3, but prefers s_2 to s_1 otherwise). Here, it is assumed that hospitals' preferences over residents are *responsive*: they always prefer to add an applicant s_i to a group of residents rather than applicant s_j (or to leaving a place empty), just in case s_i is acceptable $s_i \succ s_j$. See Roth (1985). Idealising assumptions will be subject of section 3.4.2.

[9] It is common to describe the algorithm using the predicates 'propose' and 'accept'/'reject'. Of course this refers not to the agents' behaviour in a decentralised market but to the algorithm's processing of the ROLs.

on its ROL, until its quota is filled. Each student tentatively "accepts" the highest-ranked proposer on her ROL, and rejects the other proposers.

- In the n-th step, each hospital subject to rejections in step $n-1$ proposes to the highest-ranked students to whom it has not previously proposed until its quota is filled. Each student tentatively accepts the highest-ranked hospital on her ROL among the proposers or the hospital she tentatively accepted in the previous step, and rejects the others.

- The process is repeated until there are no more proposals, at which point the students are matched to the hospitals whose offers they are holding (or remain unmatched otherwise).

As shown in the seminal article Gale and Shapley (1962), in the simple model described above, DAA^H implements *stable* matchings with respect to the ROLs submitted. A matching is stable if no one is matched to an unacceptable partner, and there is no *blocking pair*: a pair that consists of a student and a hospital that are not matched to each other but each is higher-ranked on the other's ROL than some partner assigned to them in the matching.[10]

3.2.2 Reforming the Market: Three Lessons

The concept of stability and the fact that DAA^H implements it are results from matching theory. The original algorithm wasn't designed with the help of models from this theory, and consequently, it was unlikely known to be

[10]The intuition behind the proof that DAA^H implements stability is simple: under this procedure, no one can be matched to an unacceptable partner, and there can be no blocking pair because, if a student s_j is ranked higher on a hospital h_i's ROL than a student matched to it, h_i must have applied to s_j at some previous step and been rejected. Thus s_j must have ranked h_i lower than her actual match and so (s_j, h_i) is not a blocking pair.

stable in this sense.[11] Intuitively, stability is an important concept because, assuming that agents submit ROLs that reflect their preferences, the absence of blocking pairs removes incentives for making deals outside the system. This suggested that stability was the formal equivalent to the directors' first goal to provide incentives to stick to the matchings.

However, this is a hypothesis on the basis of the model alone. To gain confidence that stability would really achieve this goal, the designers resorted to natural experiments. Among others, there were regional matching markets for physicians and surgeons in Britain. Of the eight markets investigated, six used unstable mechanisms and only two of them had survived by the time the study was made (Roth (2002)). The two remaining markets used stable algorithms, and both were performing well. This gave evidence for the importance of stability. It was still logically possible that the survival or not of the different markets was due to other factors than stability. In order to dispel this doubt, simple environments were created in laboratory experiments in which the only difference would be the mechanism in use. The experiments reproduced the field results, thus providing confidence that stability is key for achieving the first goal stated by the directors. I take this to be the first lesson from the NRMP: *a simple model suggests properties which may correspond to policy goals, and mechanisms that implement those properties. Then experiments that mirror the model – natural, laboratory, or others – provide evidence that properties "work" in the real world and can be brought about by the mechanisms.*

Stability seemed to be the directors' first-order goal. With respect to their second-order goals, the outcomes of different stable mechanisms can be com-

[11] By the time the NRMP directors felt the need for a new matching mechanism they were apparently aware of some results from matching theory. Roth recalls a personal conversation with David Gale in which the latter mentioned that he had already sent a copy of Gale and Shapley (1962) to an administrator of the NRMP in 1976. He adds that this "seems to have been the first time that anyone associated with the program became aware of the game-theoretic formulation of the problem and the results concerning [optimal stable matchings]" (Roth, 1984, 1001).

pared in the simple model above. Suppose we substitute DAA^S for DAA^H, which is the equivalent algorithm but with the roles of the students and the hospitals switched. DAA^S produces matchings that are stable too, but which are weakly preferred by all students to all other stable matchings with respect to their ROLs submitted, whereas the hospitals weakly prefer the matchings from DAA^H to all other stable matchings. So it could be hypothesised that DAA^S would perform better with respect to the second goal to produce stable matchings as favourable as possible for students. Furthermore, DAA^S makes it a dominant strategy for all the students to submit their true preferences, whereas DAA^H does not make it a dominant strategy for either side of the market to reveal their preferences (an asymmetry that stems from the fact that hospitals take multiple students whereas students are assigned to a single hospital). Moreover, there will be some room for strategic behaviour, as there is no stable algorithm that makes it a dominant strategy for all agents to reveal their preferences. Since the final goal was to reduce opportunities for strategic behaviour, and this was particularly conspicuous on the part of the students, DAA^S might be considered the algorithm of choice.

However, our simple model lacks relevant features of the market. As described above, there are couples among the applicants that are permitted to hand in ROLs specifying pairs of positions. Couples are absent in the model above, but they can be added to it. Which leads to the second lesson: *models are intervened on; different mechanisms are tested within a model, and features of the market previously missing are accommodated.*

Some of the theorems described above do not generalise to models with couples; in particular, the set of stable matchings can be empty (Roth (1984)). Such negative results pointed at some of the problems that could arise in the real-world market. The designers next asked whether there was an algorithm that would produce stable matchings whenever they exist. A simple deferred acceptance algorithm (modified to process couples' ROLs specifying pairs of

positions) would not do this job – which explains the fact that when couples entered the market in the 1960s, rates of participation dropped.[12]

Roth and Peranson (1999) designed a modified DAA^S which seeks to find stable matchings by detecting blocking pairs and repairing them, if possible, at intermediate steps. The Roth-Peranson algorithm is much more complex than a simple DAA^S (for an insightful graphical representation of the algorithm, see Roth (2013)). Many questions about its design could not be decided through existing theorems: for example, effects of different sequencings of proposals were not known. In order to compare the performance of different designs, computational experiments were made using ROLs submitted in previous years.

Another example of computational experiments will lead to the last lesson. Because the set of stable matchings can be empty, there was of course no guarantee that the Roth-Peranson algorithm would always find a stable matching. But computational experiments suggested that, under certain conditions (not too great a proportion of couples and sufficiently short ROLs), in large markets stable matchings exist with a high probability. So the result in the model located problems, and suggested computational analyses to investigate magnitudes that were, by the time, not known from the model. Interestingly, these analyses in turn prepared the ground for new theory: the computational results suggested that there could be theorems proving the existence of stable matchings in large markets. This intuition turned out to be correct about a decade later, when Kojima et al. (2013) proved analytically that, if there are sufficiently small numbers of couples and ROLs are short, as a market becomes large, the probability that a stable matching exists tends to certainty. The

[12]Roughly, the problem is the following. Suppose DAA^S is running, and the members of a couple are both tentatively accepted by two programmes. Then, if in the next step the first (but not the second) gets displaced by a preferred applicant, the couple applies to the next best preferred pair of positions which means that the second member of the couple is withdrawn from the programme that had tentatively accepted her. But then blocking pairs may occur between that programme and applicants it has rejected in order to hold the second couple member.

final lesson from the NRMP is that, *not only do (lab, field or computational) experiments provide evidence for theoretical hypotheses; they also point at new theory, in which case it is theory/analytical models that confirm results from those experiments.*

The Roth-Peranson algorithm has since its introduction in 1998 found stable matchings every year, and it also performs well with respect to other policy goals, for example, it practically makes it a dominant strategy for applicants and programmes to state their true preferences (Roth (2013)). It is a prime example of successful market design; but apparently, it doesn't fit well with accounts that treat models and empirical studies, or theorists and experimentalists, as rivals.

3.3 The Need for a General Epistemology of Economic Engineering

The reforms of the FCC and the NRMP differ in various respects. First, the kind of "goods" to be allocated partly determine the kind of market to be designed: the FCC organises auctions to allocate spectrum, whereas residents are of course not auctioned, but allocated through two-sided matchings in which the currency is preferences not money. Second, in the design of the medical match, a centralised matching system already existed, which had to be changed, whereas the spectrum auctions were to be designed from scratch. And third, the relative importance of models and various experimental methods seems to have differed in both cases. For instance, in the NRMP, field data complemented models in providing evidence that stability matters. In the auctions, where such field data were largely absent and the models available more circumscribed, experimental test beds were heavily drawn on.

A general account of how knowledge is generated in economic engineering must be consistent with both cases. A simplistic view that gives credit to

models alone cannot account for the fact that in both cases, experiments were also needed. This view thus proves inadequate, as Alexandrova and Guala convincingly show. On the other hand, accounts such as Alexandrova's, which highlight the role of experiments at the expense of models, or which generally treat theorists and experimentalists as competitors, are inadequate, especially for explaining the NRMP.

I shall argue in the following that the three lessons from the NRMP hold more generally and can be seen as a starting point for a general epistemology of economic engineering. In both cases, the initial step was to come up with a model of either the existing market (in the NRMP), or of a projected, possible market (in the FCC). For the FCC, experimentalist Plott provides evidence: "Designs are motivated by ... a mathematical model, a body of theory .. that is perhaps completely devoid of operational detail" (Plott, 1981, p. 134). Empirical methods provided important evidence in the design processes, but models came first, both chronologically and epistemically. This is not a coincidental order, but rather, I contend, the typical case. First, a model is typically needed to give precise meaning to policy goals, such as removing incentives to make deals outside the system, by defining properties, such as stability. These properties are defined relative to assumptions of the model, e.g. idealised preference structures. The properties may strictly speaking be meaningless in the real market, where those assumptions are unlikely to universally obtain. Second, a model is typically needed to provide guidance to what kinds of mechanisms should be tested. Models exclude those mechanisms that have no chance of leading to desirable properties, such as unstable algorithms. Narrowing down the potentially infinite number of mechanisms to those that could possibly implement these properties would usually be hopeless through trial and error (cf. Jackson (2018)).

The second lesson, too, holds quite generally: not only are target systems modelled, but those models are intervened on to see how outcomes change

if the model structure is altered. Even outside engineering contexts, it has been noted that this is an important function of models. Cartwright (2009) writes, "we probe models as a means to understand *how* structure affects the outcomes" (p. 57, emphasis original). According to Morgan (2012), this is even a defining characteristic of models: what makes a model a model is the fact that it can be manipulated. Manipulability is particularly perspicuous in engineering, where economic institutions are altered or created, which directs attention to the way in which counterfactual information is encoded in the models used. This will be made precise in the next section. Note, however, that it does not mean that one can naively read off what happens in possible alterations of the circumstances. In both the FCC and the NRMP, when adding complications of the real market to a model, it was found that some theorems holding in the simple model did not hold in the resulting, more complicated one. These negative results directed attention to experiments, which wouldn't have materialised without prior intervening in the models.

The final lesson from the NRMP is that, not only do models suggest hypotheses, e.g. that a mechanism implements certain properties, but sometimes experiments suggest hypotheses, which may in turn be confirmed analytically in the model. This lesson is also true of the FCC auctions, which sparked a comprehensive new body of theory. Thus, not only do models used in market design call for experiments, but the converse is also true. Roth famously stated this observation thus: "*in the service of design, experimental and computational economics are natural complements to game theory*" (Roth, 2002, p. 1342, emphasis original). I shall next provide a general account of economic engineering, which does justice to these lessons.

3.4 A General Epistemology of Economic Engineering

In line with the view that models typically "come first", I start in the next subsection by giving an account of models, which coheres with the logic of economic engineering: it shows how models project structure and allow tracking how interventions change their outcomes. In the subsequent subsection, I show that complementary uses of models and empirical methods enable inferences about interventions in the real world.

3.4.1 An Interventionist Account of Models

I shall focus on models from game theory. Both auction and matching theory consist of game-theoretic (GT) models, showing the importance of this class for economic engineering. For illustrative purposes, I will introduce a class of simple textbook, non-cooperative GT models that are not directly applicable to either case, but most of what will be said holds for more complex GT models and other economic models as well.

GT models are formal structures of sets and relations between them. Interpretations connect the formal structures to target systems. This will be subject of the next section; for now, suffice it to say that the intended interpretation is, roughly, that the sets correspond to the agents involved, their possible choices, their beliefs, desires, and reasoning processes; and the relations between them reflect how the agents' beliefs, desires and reasoning processes result in choices, as well as which outcomes result from combinations of choices.

Engineers model institutions as variables to be intervened on (cf. the classical expositions Hurwicz (1972, 1973)). Taking this view literally, I use directed graphs to describe the structure that models impose on the defined sets, as

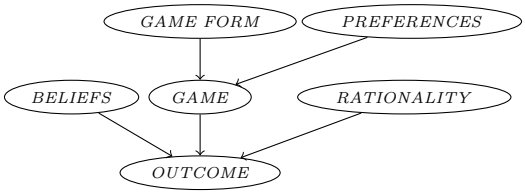

Figure 3.1: Graph of GT models. The nodes in the graph are variables and edges represent functional relations.

in figure 3.1.[13] The graph consists in a set of nodes and directed edges between the nodes. The nodes represent random variables, each of which takes values in a specific class of sets, which will be defined below. I shall talk interchangeably of nodes and variables whenever this does not cause confusion. The edges represent functional relationships, so that the value of a given node is a function of the values of its parents (that is, those nodes directly pointing at the node). For example, the graph specifies that *GAME* is a function of *GAME FORM* and *PREFERENCES*.

Informally, the graph can be described as follows. Starting from the top, the *GAME FORM* variable ranges over possible rules that govern an institution. The *PREFERENCES* variable ranges over combinations of the players' preference relations.[14] Together, these variables define a *GAME* variable whose values are particular games. The *GAME* variable, together with a *RATIONALITY* and a *BELIEFS* variable, defines the *OUTCOME* variable, which is the only leaf node in the graph. Notice that at this stage, all of these structures are purely mathematical, with no presumption about what they represent in the real world, in spite of the loaded language used in the standard definitions. For example, with "preference relation", we just mean a partial order, etc.

That the graph represents models from non-cooperative game theory (other models, e.g. from cooperative game theory, general equilibrium models, etc.,

[13] Although I apply this strategy in the context of economics, this approach might be useful in more general contexts where models express information about interventions.

[14] This node could be split into various nodes, e.g. one for each player's preference relation, but for our purposes it is convenient to summarise those relations in a single variable. The same holds for the *RATIONALITY* and *BELIEFS* variables below. Thanks to an anonymous referee for pointing this out.

may be represented by different graphs), is consistent with standard textbook presentations (e.g. Osborne and Rubinstein (1994)). These typically start by defining a game in normal form $\Gamma =< N, (A_i), (\succeq_i) >$, where N is a set of players and for each $i \in N$, A_i is a set of available actions and $\succeq_i$, her preference relations over action profiles $\mathbf{A} = \times_{i \in N} A_i$.[15] For now, think of the *GAME* variable in the graph as ranging over Γ.

The definition of a game is in textbooks usually followed by the introduction of solution concepts, mappings from games to (sets of) action profiles, which select a value for each A_i. Then, the epistemic conditions are presented that constitute solution concepts. (At this stage, this too is simply a mathematical statement about how certain conditions constrain the possible values that the A_i can take.) These constraints are represented by the *BELIEFS* and *RATIONALITY* nodes in the graph, whose values together with *GAME* define the value of *OUTCOME*, which is a set of action profiles. For example, the most prominent solution concept, Nash Equilibrium, is obtained, roughly, if the *RATIONALITY* variable takes as value that every player chooses an optimal strategy, and that the *BELIEFS* variable takes as value that players hold correct beliefs about the game they are playing (this includes complete information about the opponents' preferences) and the rationality of the opponents, and there is common belief in the players' strategy choices. Or, consider the case in which the *BELIEFS* variable takes as value not that players know the opponents' preferences but that probability distributions over preferences are commonly known. This constitutes a game of incomplete information, and the solution concept is Bayesian Nash equilibrium. In order to define players' rationality and beliefs as set-theoretic entities, epistemic models must be introduced (which would lead us too far astray, but cf. Aumann and Brandenburger (1995)). The lower part of the graph – the variables *BELIEFS*,

[15] For ease of exposition I will ignore mixed strategies. They could be added in the standard way: assume that players' preferences range over lotteries on action profiles. Then, if their preferences follow the von Neumann-Morgenstern axioms, they can be represented by payoff functions $u_i : \mathbf{A} \longrightarrow \mathbb{R}$, for all $i \in N$.

GAME, *RATIONALITY*, and *OUTCOME* – reflects the standard definition of a game plus solution concept, which result in action profiles.

A more general case is that the *OUTCOME* variable ranges not over sets of action profiles but over their consequences. This case can be accommodated if we add to the game a set of consequences C, and $g : \mathbf{A} \longrightarrow C$ a function from action profiles to consequences. The players' preferences ($\succeq_i$) are defined on C. So in this general case, the *GAME* variable is a tuple $< N, (A_i), (\succeq_i), C, g >$, and the *OUTCOME* variable ranges over C. Finally, we distinguish between two parts of the *GAME* variable: the preferences, and the remaining parts. Accordingly, two separate variables are defined: a *PREFERENCES* variable that ranges over ($\succeq_i$), and a *GAME FORM* variable that ranges over $< N, (A_i), C, g >$. I choose to present these as the exogenous variables that map into the *GAME* variable because, since we are concerned with the design of institutions, it will be important to distinguish the institutional, 'public' parts of the game which can be changed or imposed as a policy (the game form) from the 'private' parts that cannot (the players' preferences).

Remember Cartwright (2009)'s claim that models are examined to discover how their structure affects outcomes. Within the graphical representation of GT models, in which their constituents are treated as variables, this intuition can be made precise: it amounts to model-interventions. The following example uses a simple game to illustrate these interventions.

Example 3.1 (Prisoners' Dilemma and Prisoners' Delight) *Suppose GAME FORM takes as value the rule specified in table 3.1. There are two players, Row and Col, who can choose to cooperate or defect. The four possible action profiles result in consequences a, b, c, and d.*

We assume that $PREFERENCES_{Row} = c \succ a \succ d \succ b$, *and* $PREFERENCES_{Col} = b \succ a \succ d \succ c$. *This determines the value of the GAME variable: we have a Prisoners' Dilemma. Suppose RATIONALITY takes as value that play-*

GAME FORM=		cooperate	defect
	cooperate	a	b
	defect	c	d

Table 3.1: Value of *GAME FORM*: rule governing a two-player interaction between Row and Col that can both choose to cooperate or defect. The action profiles result in outcomes a, b, c, or d, as shown in the table. When a value is specified for *PREFERENCES*, this defines the value of *GAME*.

ers play optimal strategies, and BELIEFS, that they hold correct beliefs about the game. (Note that beliefs about the opponent's rationality don't matter in this case because both players have a strictly dominant strategy.) Then, both players will defect and $OUTCOME = d$.

GAME FORM=		cooperate	defect
	cooperate	a	c
	defect	b	d

Table 3.2: An intervention on the *GAME FORM* variable: switching the asymmetric outcomes.

Suppose we intervene on the GAME FORM variable, which results in the value specified in table 3.2. This illustrates a change of the institutional rules, in which the asymmetric outcomes are switched. It induces a different game ("Prisoners' Delight") in which cooperation is a dominant strategy for both players. If we suppose that RATIONALITY and BELIEFS are held fixed at the values of before, then $OUTCOME = a$*, which is the second-best as opposed to the second-worst outcome for both players.*

The example shows that game theory encodes counterfactual information in natural collections of its models: it allows calculating what the corresponding outcomes are for different values of the variables. The counterfactual information encoded in a model is at the core of my argument of how game theory, or economic theory more generally, can be harnessed for economic engineering. Engineers use this information to intervene on the *GAME FORM* variable: game forms are designed which "force" players to produce outcomes that are considered desirable.

3.4.2 From Model-Interventions to Real World-Interventions: the Complementarity of Models and Other Tools

Within a GT model, different games can be compared and the one selected whose equilibria come closest to the policy goals in question, or more precisely, to what are interpreted to be the policy goals within the model. The idea is to make the market resemble that game, in particular by imposing rules that correspond to the game form in question. If the design is successful, those rules induce the intended policy goals in the real world.

This requires a kind of *external validity*: roughly, that interventions within the model inform interventions in the real market in the sense that the latter interventions reliably establish outcomes that resemble those in the model.[16] So it seems that what is required is the truth of a causal hypothesis: viz., that there exists a causal mechanism in the real world, to which the structure of the model corresponds; and that the game form in the model implements properties that correspond to the relevant policy goals. The truth of this hypothesis would justify confidence in the prediction that institutional rules corresponding to the chosen game form in the model, will produce desirable outcomes.

The hypothesis can be made precise in terms of our graphical representation: what it says is that the directed graph of a given model can be interpreted as a *causal graph* of its target system (which may be a possible, future system). A causal interpretation of a graph consists in, first, specifying that its variables range over events, and second, that its edges correspond to causal

[16] "External validity" usually refers to the generalisation of laboratory results to circumstances outside the lab. For a general methodology of external validity in economics, cf. Part II of Guala (2005). Here, I use external validity in a broad sense that includes inferences from models to the real world.

relationships.[17] As an example of the first, suppose the graph in figure 3.1 is a causal graph of a particular interaction. Then the variable *GAME FORM* takes as value the event that a specific rule governs that interaction, specifying which choices are available to agents and how their choices jointly result in outcomes. The *PREFERENCES* variable takes as value the event that agents have specific preference relations. The *BELIEFS* variable takes as value the event that agents have specific beliefs about the situation they are in, and their opponents' reasoning processes. And so on. Second, its directed edges denote *direct causation*:[18] roughly, that there are different values of the parent node such that changing it from one to another, the child node changes its value from one to another, given that its other parent nodes are held fixed at specific values.

If a graph can be interpreted causally, the structural relations in the model to which it corresponds describe a causal mechanism effective in the real world. Model interventions can then reveal information about how the outcome of the target interaction would change as the rules that govern it change. However, economic models famously make false assumptions. False assumptions pose a problem for the external validity of model interventions because they seem to prevent a causal interpretation of their associated graphs. These graphs may be *imperfect*, that is, the values of some variables are not known, or it is known that values specified in the model are false of the target interaction. For example, standard GT models assume that agents are perfectly rational, but they may not be in the projected interaction. But this assumption is needed to identify equilibria. So we lack confidence that agents act according to our predictions based on equilibria. Graphs may also be *incomplete*, that is,

[17] To show that it is in principle possible to interpret a graph causally in the sense of, e.g. Spirtes et al. (2000), requires specifying a probability distribution over the graph which satisfies the causal Markov condition and the minimality condition. Under the standard GT textbook interpretation of the nodes that I gave above, the graph in figure 3.1 trivially satisfies these axioms, thus allowing for a causal interpretation.

[18] I do not wish to commit to a specific theory of causation. It should be clear though that theories in which interventions figure prominently – such as Woodward (2003)'s theory – fit well with my account of models.

lacking nodes or edges that may change the outcomes of target interactions. For example, recall the winner's-curse phenomenon in auctions, in which the bidder who most overvalues an item wins. Models from auction theory suggested that open auctions could reduce the risk of the winner's curse, but it was not clear whether they would do so in the presence of complementarities, which did not figure in the models.

Typically, economic models (or rather, their associated graphs) are both imperfect and incomplete. How can the external validity of such models be established? In some cases, suitably constructing the model and intervening on it goes some way towards it. A prime example is the treatment of private information. In order to know what game is induced by a game form, the value of the *PREFERENCES* variable (called the players' *types*) must normally be specified. But the designer does not know the players' types, and since they may have incentives to hide their preferences, it is of no help to simply ask them. Indeed, one of the problems encountered in our case study was precisely that agents were trying to "game the system" by handing in ROLs strategically.

Theorists seek to overcome this problem by designing "robust" game forms, that is, game forms that implement desirable properties for variable types and populations of players (cf. Kuorikoski and Lehtinen (2009) for the logic of robustness). For instance, deferred acceptance algorithms implement stability in the simple model above with respect to ROLs submitted, no matter whether a given agent prefers Chicago to L.A., or the other way around. So the imperfection of the model that arises from the lack of knowledge of the players' preferences is to an extent outwitted. "To an extent", because doing so requires assumptions on the agents' preferences (e.g. that they are complete and transitive) and agents must be given incentives to reveal their preferences, otherwise there may be instabilities. But, as we have seen, there is no mechanism that achieves this for all agents.

At this stage, complementing models with empirical data may establish external validity. In the NRMP, natural experiments gave evidence for the well-functioning of matching markets that use stable algorithms and the malfunctioning of markets that do not. These findings could be replicated in the lab, where the stability property was identified as the main cause of the well-functioning of simple matching markets. By showing that model results can, or cannot, be replicated in lab environments, experiments confirm, or correct imperfect graphs. Experimental systems are also used to supplement incomplete graphs by introducing mechanisms which the model is silent about, as in the noted, holistic treatment of experimental test beds. All these cases vindicate the view that experiments "mediate", or "bridge the gap", between models and their intended target systems (cf. Guala (2005); Guala and Mittone (2005)). As we have seen, empirical data may also lead to new theory. In the NRMP, this took the form of running the algorithm over sample data, which suggested a large market theorem ("in a market with couples, the set of stable matchings is unlikely to be empty if the number of participants is large...") that could later be proven analytically.

What combination of tools – models, laboratory or field experiments, computational methods, etc. – achieves reliable inferences, and which tools figure most prominently in a design process, may differ from case to case. The differences between the NRMP, where models went a long way towards the final design, and the FCC, where their applicability was more circumscribed, are a case in point. A recent proposal by Esther Duflo can be used to put different design processes in rough order. Duflo (2017) introduced the term "plumbing" for cases of economic design in which it is uncertain what the relevant features of a target system are. So, for example, the NRMP would have been a case of plumbing had there been no data available confirming that stability causes a market's well-functioning. There is no single, clear-cut feature that distinguishes plumbing from engineering; rather there is a spectrum, on which plumbing lies on one end, abstract mechanism design on the other, and

engineering occupies an interval in between. Moving towards the plumbing end of this spectrum, things get more detail-focussed, the knowledge created more context-specific, and models can be relied on to a lesser extent. It seems fair to say that, while both the NRMP and the FCC are cases of economic engineering, the latter is closer to the plumbing end of the spectrum.

Where on the spectrum a given case is located depends on the specific characteristics of the target system and on the existing models and other tools that the designers can resort to. It is not usually known ex ante exactly where on the spectrum a given case will be located. Ex post, we may be able to reconstruct design processes (as I have done for the NRMP) and to give an account of what combination of tools got the designers ahead. But ex ante, the economists' intuitions, and to an extent trial and error, determine the methods used for the design. I conclude by considering the implications of this insight for efficiency analysis.

3.5 Conclusion: Economic Engineering and the Efficiency Question

What combination of models, experiments, or other tools, provides efficient ways to design markets? An answer to this question, either globally or for a given case, has implications for how resources should be spent. Northcott (2018) rightly points out that efficiency analyses are inevitable and happening anyways, such as when researchers decide to model an institution, or to organise experiments, or when teams of theorists or experimentalists are assembled. But Northcott urges philosophers of science to explicitly keep tabs on current practices: to analyse and assess whether they are efficient, or whether a shift of resources is commendable, for example towards experiments and away from models.

The NRMP and the FCC are prime examples of successful economic engineer-

ing. Contrasting their design processes, we found that the relative importance of different tools differed in the two cases and that it is not usually known ex ante what methods will move us forward. This urges some caution, at least on a global approach to efficiency analysis. Juggling the variety of different cases on the spectrum of economic design in global efficiency claims may be seriously misleading and might thwart the goals of efficiency analysis.

Would the design process of the NRMP have been more efficient had a different combination of methods been applied? This is what the local efficiency question for the NRMP amounts to (equivalently for the FCC). It could mean either that the matching mechanism, which materialised and which turned out to be successful, could have been designed using fewer resources had a different combination of methods been applied; or that the same resources, allocated differently, could have led to a design at least as successful as the default. Both counterfactuals are, I believe, practically impossible to assess. Fewer resources spent would have led to different teams of researchers, to different combinations of models and experiments used, and so on. In other words, they would have led to a different history of the NRMP, and a fortiori would likely have brought about an altogether different design. Because of the complex and potentially long causal histories involved, it is hard to assess whether the same institution could have been designed more cheaply. It is equally hard to evaluate whether the same resources, allocated differently, could have achieved a more successful design. This would require assessing what that design would have looked like, which again involves complex and potentially long causal histories.[19]

This problem is inherent to engineering. It may be straightforward to ask for a given explanandum whether one or another explanans (which may use different mixes of models and other tools) performs its task of explaining

[19] A futher difficulty is, when does the relevant causal history start in the first place? For example, the model introduced in section 3.2 traces back to Gale and Shapley (1962)'s marriage model. Should the causal history be traced back to 1962?

more efficiently. For example, Northcott compares Axelrod's game-theoretic analysis of truces in World War I trench warfare to historical analyses of the same phenomena. But when a market is designed, it is the result of a particular design process – a causal history that includes policy makers setting policy goals, the teams of researchers and the techniques they apply to project causal structure and intervene on it. Had this history been different, the product would likely be different.

I take these to be reasons to abstain from philosophical efficiency analysis, at least in the context of economic engineering, and to promote free research instead. Paul Milgrom, one of the protagonists of the FCC auctions, writes, "We are celebrating the fruits of research that could just as easily have found itself ridiculed. Who knows what tomorrow will bring?" (Milgrom (2017)).

Chapter 4

Towards a Fair Distribution Mechanism for Asylum

Recently there has been increasing interest not only in the number of refugees that countries (should) accept, but also in achieving good "matches" between refugees and their host countries. For example, the mechanism for relocating refugees from Greece and Italy to other member states of the European Union (EU) seeks to realise this goal by allowing countries offering relocation to indicate preferences over refugees (see European Commission (2015)). Good matches are important because a refugee's international protection needs, and her opportunities to flourish, are served differently in different countries. Moreover, a country's costs for hosting refugees and the public opinion towards them may differ for different types of refugees, which may affect policy makers' willingness to comply with international legal norms (Bansak et al. (2016)).

The question of how to distribute asylum amounts to a problem of designing refugee distribution mechanisms according to criteria that may be considered desirable or morally required. Relevant criteria may include maximising the number of places for refugees, fairness, or efficiency considerations. The aim of this chapter is to provide some insights into the normative issues that the

design of distribution mechanisms raises in this context. One important issue is this: satisfying refugees' or countries' preferences may in some cases reduce the number of refugees matched. It is argued here that there is no simple solution to this problem, and that in instances where a trade-off between the satisfaction of preferences and the number of refugees matched occurs, it is not a reasonable policy to take preferences into account.

On the positive side, I show that a simple sufficient condition can be given which, if satisfied, precludes the trade-off from occurring: that all the countries within the system deem all refugees acceptable, and that all refugees deem all countries acceptable. The latter can only be required if each refugee's rights are respected in each country within the system, rather than in the country she is matched to alone. I interpret this as a precondition that must be satisfied for an asylum policy to reasonably take preferences into account.

Finally, I argue that, in an appropriate decision framework for asylum, countries should not be allowed to express preferences over groups of refugees in the first place. Instead, priorities over refugees should be imposed that take into account humanitarian factors such as vulnerability, as well as fairness conditions among the countries within the system. Furthermore, in the context of distributing asylum, the elimination of justified grievance is arguably a weightier normative criterion than efficiency considerations.

In order to make the arguments precise, I draw on tools from game theory. The distribution of asylum is modelled as a (bipartite, many-to-one) *matching problem under preferences*: refugees and countries offering asylum make up a two-sided market in which the members of one side are to be distributed over members of the other side. Moreover, members of the market have preferences over or may give priority to members of the other side of the market, the satisfaction of which makes for the goodness of the matchings. Formulating the asylum market in this way allows us to use tools from matching theory to provide a more precise understanding, and may eventually contribute to

the implementation of more fair or efficient policies. Various social scientists have recently argued for imposing centralised matching systems and have investigated mechanisms that could be implemented in different stages of asylum seekers' path to their final destination (Fernández-Huertas Moraga and Rapoport (2014, 2015a,b); Jones and Teytelboym (2017a,b, 2018); Delacrétaz et al. (2016 version); Andersson and Ehlers (2018)).[1] While I am sympathetic to this line of research, this chapter does not directly contribute to it. Rather, it serves as a commentary on normative issues that typically arise in this context.

The remainder of the chapter is organised as follows. In the next Section, our main case study is introduced: the relocation mechanism that is currently in effect in the EU. The case study naturally suggests three desiderata on the distribution of refugees that are made precise in Section 4.2 through a simple model from matching theory (the "College Admissions model"). A mechanism that takes preferences into account is shown to achieve some of the desiderata. I also show that there may be trade-offs between two of the desiderata, and so a condition is given to prevent these trade-offs from occurring. However, in Section 4.3, it is shown that within this model one of the desiderata is violated: the model opens the door to discriminatory policies and unduly favours popular host countries at the expense of less popular ones. This motivates designing an asylum market that corresponds to a different model (the "School Choice model"). This model is described in Section 4.4, where I also argue that fairness beats efficiency in the context of distributing asylum. Section 4.5 concludes.

[1] See also the organisation *Refugees' Say* (`https://www.refugees-say.com/`), which aims to "empower refugees and communities" by developing resettlement schemes that account for their preferences.

4.1 Desiderata on the Distribution of Refugees: The EU Relocation Mechanism

To prevent terminological confusion, let *refugee* denote a recognised refugee or an asylum seeker with a justified claim to refugee status.[2] We are concerned with their distribution on a supranational scale: refugees are to be distributed over a given set of nation states (henceforth 'countries') which—voluntarily or enforced by a superordinate (con)federation government—agree to be possible host destinations. The distribution problem may either occur in the context of *resettlements*: the distribution of refugees from third countries or refugee camps among the countries in the system; or in the context of *relocations*: the redistribution of refugees already in a country in the system.

The context determines the number of refugees in the system. This can be a target set by a superordinate institution (such as the EU's target to relocate a certain number of refugees), or the sum of pledges made by the countries within the system (such as some EU member states' pledges to resettle a certain number of refugees). Moreover, there may be (although there need not be) quotas: numbers of refugees that countries will accept. These numbers may be individual pledges made by the countries; or they are imposed according to a distribution key (e.g. as proposed in Bartsch and Bovens (2016) or Grech (2016)); or they may be the outcome of a market of tradeable immigration quotas (as suggested in Fernández-Huertas Moraga and Rapoport (2014)).

Let us take a concrete distribution problem as a case study. In September 2015, the European Commission proposed a mechanism for the relocation of refugees from Greece and Italy among EU member states. The proposal shows awareness of the importance of good "matches" between refugees and countries of relocation:

[2]This presupposes a consistent definition of what counts as a justified claim (at least among the countries within the system), which is a controversial issue (see Shacknove (1985)). For simplicity, this problem will be ignored except for a short discussion in Section 4.4.

> "[I]n order to decide which specific Member State should be the Member State of relocation, a specific account should be given to the specific qualifications and characteristics of the applicants concerned, such as their language skills and other individual indications based on demonstrated family, cultural or social ties that could facilitate their integration into the Member State of relocation." (European Commission (2015))

Subsequently, the Council adopted the proposal in a decision on a temporary relocation of 160,000 asylum seekers in clear need of international protection from Greece and Italy (Council of the European Union (2015, 2016)). In order to comply with the target of achieving good matches, the relocation mechanism allows member states of relocation to indicate preferences over refugees who applied for relocation. Greek and Italian authorities then choose among applicants and thereby "try as much as possible to meet the preferences expressed" (Council of the European Union (2016)). After determining the matches, they send relocation requests to the countries, which are legally binding.

However, the relocations are not running smoothly. Three problems are particularly noticeable. First, there seems to be considerable discontent with some matches. Many refugees have disappeared after learning about the decision on their destination country (European Commission (2016a)). Others have vanished after their relocation, and preventing such irregular secondary movements has become a central policy goal (Council of the European Union (2015); European Commission (2016c)). Second, the number of refugees distributed lags far behind the policy target. By July 2016, almost a year after the Commission's proposal had been adopted, the total number of persons relocated equalled only 3056, which corresponds to less than 2% of the 160,000 people envisaged (European Commission (2016b)). As of July 2017, the mechanism seems to have gained some traction, but still less than 25,000 refugees were relocated (European Commission (2017)). Third, the preferences that some

countries express are ethically problematic and in conflict with the EU's policy goals. For example, although the Council appealed to the member states to prioritise particularly vulnerable persons (e.g., unaccompanied minors, pregnant women, disabled and elderly persons), some member states are reluctant to receive persons from these groups (European Commission (2016a)). Furthermore, although legally required to accept all types of refugees (including those low-ranked on their preference orderings), some member states have rejected allocations on the grounds that their preferences were not respected (European Commission (2016a)).

The problems with the EU relocation mechanism motivate three desiderata on asylum matchings that will be adopted in this discussion. First, available places should be used efficiently (in a sense to be specified). Roughly, the refugees' or countries' preferences should be satisfied "as much as possible". In particular, refugees' incentives to vanish or partake in secondary movements should be minimised. Second, the number of refugees matched should be maximised and should possibly equal the policy goal. Third, the system should ban what is called an "incorrect use of preferences" (European Commission (2016a)): preferences should be expressed in line with fundamental ethical principles and higher-order policy goals.

The desiderata can be defined precisely within the framework of matching theory, which shall be introduced next.

4.2 Asylum as College Admissions Problem

Matching under preferences is a tool from cooperative game theory. It can be applied to two-sided markets in which heterogeneous agents, or goods, of one side are to be distributed over agents or goods of the other side of the market, and the satisfaction of agents' preferences, or respect for agents' priorities, matter. Gale and Shapley (1962) laid the theoretical foundations

for the theory. Centralised matching systems have since been implemented in different contexts such as matching students to universities, job-seekers to employment, or kidney donors to patients.

Various economists and political scientists have argued that implementing a matching system in the context of asylum would be beneficial for refugees, or their possible host countries. Jesús Fernández-Huertas Moraga and Hillel Rapoport were the first to suggest a matching system which is embedded in the tradeable immigration quotas system they propose for, among other things, the distribution of refugees over EU countries (see Fernández-Huertas Moraga and Rapoport (2014, 2015a,b)). Moreover, Will Jones and Alex Teytelboym advertise the implementation of matching systems both on a global and on a local scale (see Jones and Teytelboym (2017a) and Jones and Teytelboym (2018), respectively). David Delacrétaz, Scott Duke Kominers and Teytelboym propose specific mechanisms for locally matching refugees with communities in different institutional and informational settings (Delacrétaz et al. (2016 version)); and Tommy Andersson and Lars Ehlers design a matching system in the context of assigning private housing to refugees in Sweden (Andersson and Ehlers (2018)).

Rather than argue for a specific matching system, or matching systems generally, our interest here is to show that matching theory yields insights that are important for the ethics of asylum distribution. The main difference between the EU's relocation mechanism and a centralised matching system is that, in the latter, the matchings are determined through the application of a mechanical procedure. The resulting matchings can be compared along the same properties, however, and thus the desiderata within the framework of matching theory and the theoretical results hold equally for contexts such as the relocations in the EU.

Fernández-Huertas Moraga and Rapoport (2014) propose implementing a College Admissions (CA) model. This model resembles our case study in that it

takes into account the preferences countries have over (groups of) refugees, which is why it serves as a natural starting point. It differs from the case study in that members of either side of the market are equally treated as agents with preferences over members of the other side of the market: countries have preferences over refugees and refugees have preferences over countries, which is why the CA model allows for more general mechanisms than the EU's relocation mechanism which only takes countries' preferences into account. A matching affects the countries' and refugees' welfare relative to the satisfaction of their preferences.

Formally,[3] a CA-instance of a refugee-country matching problem is a four-tuple $(C, R, q, \mathbf{P})$, where $C = \{c_1, \ldots, c_m\}$ and $R = \{r_1, \ldots, r_n\}$ are disjoint sets of m countries and n refugees, respectively. The agents of the market are the members $a_k \in R \cup C$. We are concerned with *many-to-one matchings* since it can be assumed that $n \gg m$ and each refugee can obtain asylum in at most one country, whereas a given country can accept many refugees. The maximum number of refugees that can be matched to each country is determined by a vector of quotas $q = (q_j)_{j \in \{1,\ldots,m\}} \in \mathbb{N}^m$. As described in the previous section, quotas may be imposed according to a distribution key, the outcome of a market of tradeable immigration quotas, or they may be individually set by the countries. But quotas might be rejected altogether on ethical grounds. The model does not take up a stance on this because it does not commit us to effective quotas. For instance, setting $q_j = n$ for all $c_j \in C$ makes them dummies. Finally, $\mathbf{P} = \{P(c_1), \ldots, P(c_m), P(r_1), \ldots, P(r_n)\}$ is a set of preference lists which induces a *complete*, *transitive*, and *irreflexive* preference profile for each country over the set of refugees and for each refugee over the set of countries. Write $c_1 \succ_{r_i} c_2$ to denote that r_i prefers c_1 to c_2, and equivalently for countries' preferences.

For the time being, we suppose that refugees may declare countries unaccept-

[3] My notation loosely follows Klaus et al. (2016).

able, and countries may declare refugees unacceptable (we shall discuss and restrict this condition later on). Hence, there is a subset $E \subseteq R \times C$ of acceptable refugee-country pairs. Denote $A(r_i) = \{c_j | (r_i, c_j) \in E\}$ the set of acceptable countries for a given $r_i \in R$; and equivalently for the countries.

An *assignment* M is a subset of E, and the set of assignees for a given $a_k \in R \cup C$ is denoted $M(a_k)$. A refugee r_i can be unassigned so $M(r_i) = \emptyset$, or otherwise assigned. Similarly, a country c_j is undersubscribed if $|M(c_j)| < q_j$, and full if $|M(c_j)| = q_j$.

Definition 4.1 (Matching) *A matching is an assignment with*

(i) $|M(r_i)| \leq 1$ *for all* $r_i \in R$*; and*

(ii) $|M(c_j)| \leq q_j$ *for all* $c_j \in C$*.*

Condition (i) says that a given refugee is either assigned to a single country or unassigned under a matching. As usual in the literature, I will use $M(r_i)$ to refer to the country to which r_i is matched instead of the singleton containing that country, whenever this does not cause confusion. Condition (ii) says that a given country accepts a subset of the set of refugees, the cardinality of which is restricted by that country's quota. In the following, we say equivalently that r_i is matched to c_j and that c_j is matched to r_i under M if $(r_i, c_j) \in M$.

4.2.1 Stability and Deferred Acceptance Algorithms

We can now turn to the desiderata encountered in the previous section. I will argue that the desideratum that preferences be satisfied "as much as possible" amounts to *stability* (ST) in the CA model. We shall first define this property and introduce an algorithm that produces stable matchings (that is, matchings that satisfy ST) for every CA-instance.

A matching M is *blocked* by a refugee-country pair $(r_i, c_j) \in E \setminus M$ if r_i is unassigned in M or prefers c_j to $M(r_i)$, and at the same time c_j is undersubscribed in M or prefers r_i to a member of $M(c_j)$. A matching is stable if it is not blocked by any refugee-country pair.

Deferred acceptance algorithms produce stable matchings in every CA-instance (Gale and Shapley (1962)); in the following, we call such mechanisms *stable* too. Consider the following, "country-proposing deferred acceptance algorithm", μ^C:

- In the first step, each country proposes to its most preferred acceptable refugees until its quota is filled. Each refugee tentatively accepts her most preferred country among the acceptable proposers and rejects the other proposers.
- In the second step, each undersubscribed country proposes to the next best preferred acceptable refugees to whom it has not yet proposed until its quota is filled. Each refugee tentatively accepts her most preferred country among the acceptable proposers and the country she tentatively accepted in the previous step, and rejects the other proposers.
- The process is repeated until there are no more proposals.[4]

The ending condition applies when either all refugees are matched, or all countries are full, or there are unmatched refugees and undersubscribed countries

[4]I should add two qualifications. First, μ^C is a simple algorithm which is not fit for purpose. For example, it cannot accommodate the fact that many refugees flee as couples or in families that should not be separated. It is nevertheless introduced in order to make clear the ethical problem that the size and the "quality" of matchings can be in conflict—a problem which is present in more complex algorithms (e.g., Delacrétaz et al. (2016 version)). Second, μ^C is not the only stable mechanism. The reason it is presented here is that it is the deferred acceptance algorithm which is arguably closest to implementing the mechanism used in the EU relocations in which the countries "pick" refugees according to their preferences. But whereas μ^C also respects the refugees' preferences, they are not taken into consideration in the EU mechanism. Whether a country- or a refugee-proposing deferred acceptance algorithm is preferable is debatable as both have pros and cons (for example, Fernández-Huertas Moraga and Rapoport (2014) propose μ^C, whereas Jones and Teytelboym (2017a) prefers a refugee-proposing algorithm). For our purposes, nothing hinges on this question and our arguments in the following hold equally for other deferred acceptance algorithms.

but all such agents are deemed unacceptable by the remaining partners they find acceptable. For an illustration of the algorithm, consider a small CA-instance.

Example 4.1 *There are three refugees, r_1, r_2, r_3, and two countries c_1, c_2 with $q_1 = 2$ and $q_2 = 1$. The preference relations are as specified in Table 4.1.*

Countries	Refugees
$r_1 \succ_{c_1} r_2 \succ_{c_1} r_3$	$c_2 \succ c_1$ for both r_1 and r_2
$r_2 \succ_{c_2} r_1 \succ_{c_2} r_3$	r_3 declares only c_2 acceptable

Table 4.1: Table specifying refugees' and countries' preferences. $a \succ_c b$ denotes that c strictly prefers a to b.

Apply μ^C. In the first step, c_1 proposes to r_1 and r_2, and c_2 proposes to r_2. r_2 tentatively accepts c_2 and rejects c_1, and r_1 tentatively accepts c_1. c_1 has a free place. In the second step, c_1 proposes to r_3 and r_3 rejects. c_1 has a free place but no more refugees to propose to so the algorithm stops. The resulting matching is $((c_1, r_1), (c_2, r_2))$.

It can easily be checked that the matching is stable. For example, r_1 prefers c_2 to her actual match, c_1. However, c_2 does not form a blocking pair with r_1 because it is full and prefers its actual match to r_1. Similarly for the unmatched r_3.

Why does (ST) in the CA model amount to satisfying preferences "as much as possible"? First, it implies Pareto efficiency in the CA model: agents could only be made better off by making other agents worse off. (It can easily be verified that this is true in the above example.)

Second, a distribution that satisfies (ST) can be considered to be *fair* in the following sense. Only if c_2 had free places available or preferred r_1 to its actual match would r_1 have a justified claim to be matched to c_2, but this is ruled out because the matching is stable. This lack of justified claims is what

makes for a fair distribution. Note that this condition may not be sufficient for eliminating all discontent with the matchings: it does not imply that all agents get what they want most (which is usually impossible, see Example 1). It only implies that they are matched to an acceptable partner and that they are not matched to a less preferred partner when a more preferred partner would be available.

Third, the sense of fairness that (ST) conveys can be expected to contribute to the thickness of the market: it gives agents on both sides incentives to participate in the system. Conversely, if (ST) is violated, both sides of the market may be dissatisfied. There is ample evidence that in many contexts, agents seek to arrange bilateral arrangements outside the system when this happens (we encountered this unravelling of the market in the context of matching doctors with hospitals in Chapter 3). Note that the EU relocation mechanism violates (ST) in a specific way: it does not systematically take refugees' preferences into account, which is why the resulting matchings may be considered unfair for refugees. In this context, it may be difficult for refugees to arrange bilateral arrangements outside the system. Instead, the failure to take their preferences into account may contribute to the finding from the previous section that many vanish after learning about their destination countries, or partake in illegal secondary movements.

4.2.2 Maximum Cardinality vs. Stability

Let us now consider the second desideratum encountered in Section 4.1. Call the number of refugees assigned in a matching its cardinality. The cardinality of a matching depends on the set of acceptable refugee–country pairs, that is,

the set $E \subseteq R \times C$.[5] Call *maximum cardinality* (MC) the desideratum that matchings should not waste places. More precisely, for a given CA-instance, a matching M satisfies (MC) if and only if $|M| \geq |M'|$ for all matchings M'.

In example 1 above, $((c_1, r_1), (c_1, r_2), (c_2, r_3))$ is the unique maximum cardinality matching. It is not stable because (c_2, r_1) and (c_2, r_2) are blocking pairs: both r_1 and r_2 prefer c_2 to c_1, and at the same time c_2 prefers both r_1 and r_2 to r_3. In contrast, remember that applying μ^C produced the stable matching $((c_1, r_1), (c_2, r_2))$, leaving r_3 unmatched and c_1 undersubscribed. Thus, satisfying (ST) comes at the price of failing to satisfy (MC).

The fact that (MC) and (ST) may conflict poses a problem because both (MC) and (ST) have normative appeal. In many contexts where matching theory is applied, (ST) is the primary policy goal, and there is some loss in the size of the matchings allowed for because it can be compensated through different instruments from market design. Two such instruments shall be discussed briefly and shown to be inadequate in the context of the asylum market. This suggests that there is no simple resolution to the problem.

First, in some matching markets (e.g. for medical residents, or for graduate economists' academic jobs), so-called "scrambles" have been established for unmatched and undersubscribed agents. These are decentralised post-match markets where available agents of both sides of the market can find each other and positions can be filled (Coles et al. (2010)). However, the fundamental difference between matching asylum and contexts where matching theory is usually applied is that in those contexts, both sides of the market have incentives to fill available places. In the context of asylum, many countries have

[5] The cardinality of matchings also crucially depends on countries' quotas which are taken as exogenous variables here. In passing, note that in contexts in which countries state voluntary quotas the application of deferred acceptance algorithms is problematic because they generate incentives to *capacity-manipulate*: countries may gain by stating smaller quotas. As shown in Sönmez (1997), there is no stable mechanism that is immune to capacity-manipulation. Hence, μ^C combined with voluntary quotas would incentivise countries to enter a race of diminishing their stated capacities—which is an extremely undesirable consequence for a good in short supply, and may serve as an argument against voluntary quotas.

opposite incentives, and few countries not filling their quotas would advertise this in a scramble. A scramble is thus unlikely to be efficient or even to emerge in the first place.

Second, countries unable to fill their quotas may be penalised.[6] Penalties are often difficult to impose, however, even when countries' participation in a relocation mechanism is obligatory. In the EU context, for example, it has been proposed that countries not filling their quotas pay a fine of 250,000 Euros for every assigned place that remains empty. What the prospects are for this proposal is questionable, however, as various countries, most notably members of the Visegrád Group, are virtually boycotting the relocations. Moreover, Hungary and Slovakia took legal steps against the relocations and refused to accept any more refugees before a verdict would be announced.[7] The European Court of Justice dismissed the suit,[8], but Hungary's prime minister Viktor Orbán declared that this won't change Hungary's policy of not participating in relocations.[9] While it may be possible to penalise Hungary for not complying with the EU regulations in the future, the process can be expected to be long and politically tedious. Update from September 2018: the EU parliament voted to punish Hungary over 'breaches of core values', which opens the possibility to suspend Hungary's voting rights. However, a suspension is unlikely to happen because it would require an unanimous vote, but other countries of the Visegrád Group are likely to vote against a suspension.

Besides worries about the prospects for implementing penalties, note that they make the system manipulable: it may pay for refugees to be "picky". This is the case when the revenues from the penalties are used to provide more asylum places elsewhere, thus in the best case satisfying (MC). In example 1, applying μ^C, r_1 gets assigned to her least preferred country, c_1. Now, suppose

[6] Penalties are proposed in Fernández-Huertas Moraga and Rapoport (2014) in the context of a market of tradeable immigration quotas, where the penalty is a function of the difference of the quota negotiated and the number of refugees assigned to the country under μ^C.

[7] See Rettman (2015) online, accessed on 22 September 2017.

[8] See Court of Justice of the European Union (2017), accessed on 22 September 2017.

[9] See Office (2017), accessed on 22 September 2017.

that for each place that remains empty, c_1 is penalised and with the help of the penalty, a place is created elsewhere. If r_1 prefers a place elsewhere to c_1, she would gain by declaring c_1 unacceptable even though she finds it acceptable. Although these are not definite reasons for the intractability of imposing penalties, they do suggest some caution: imposing penalties may be difficult, and may lead to undesirable incentive structures. Since these issues do not seem to have a definite solution at present, penalties will be neglected in the following.[10]

Whenever (ST) and (MC) are in conflict, we must bite the bullet and give up one of the desiderata. Which one? Consider again example 1. Suppose r_3 is justified in declaring c_1 unacceptable. For example, she may belong to an ethnic or religious group that is persecuted in c_1 (and suppose r_1 and r_2 do not belong to such a group). Refugee r_3 is then in a particularly disadvantaged situation because she cannot expect protection in one of the countries within the system. By assigning her to the acceptable country c_2, the maximum matching gives priority to r_3. Since priority is given to the worst-off, (MC) could be interpreted as a prioritarian condition.[11] It is moreover a condition that the Rawlsian maximin principle would embrace (Rawls (1971)).

Giving priority to the worst-off may come at the cost of making other agents in the market worse off. In the example, r_2 and c_2 are worse-off in the maximum matching than in the stable matching. As we have seen, (ST) implies Pareto

[10] For more arguments against penalties, cf. Jones and Teytelboym (2017a). A third resort to fix the problem is to internalise it in the matching system by imposing minimum quotas. This strategy is not relevant here because the context of asylum is different to the setting where minimum quotas are usually investigated. For example, Fragiadakis et al. (2015) investigate minimum quotas in the context of school choice. They assume that all schools are acceptable to all students and vice versa, and look at the case where the number of students is strictly between the sum of the schools' minimum quotas and the sum of the schools' maximum quotas. In the case where the number of students exceeds the number of places available—which is to be expected in the context of asylum—minimum quotas are dummies in this setting. In contrast, the problem we are concerned with is the case in which places may be wasted due to the size of the set of acceptable refugee-country pairs. The special case where students may declare schools unacceptable is considered in Fragiadakis et al. (2015), but their mechanisms allow violating minimum quotas and don't satisfy (MC).

[11] E.g. Parfit (1997, 2012).

efficiency in the CA model. Moreover, we have interpreted this property as the desideratum that agents' preferences should be satisfied "as much as possible" in the CA model. Thus, in contrast to (MC), (ST) in the CA model could be approximately characterised as "utilitarian".[12]

Which condition should be the dominant consideration? Arguably, in the context of asylum in which by definition, people are in disadvantaged and vulnerable situations, it is a more reasonable policy to give priority to the worst-off. There is empirical evidence that this agrees with a widespread intuition in many receiving countries (Bansak et al. (2016)). This suggests that satisfying preferences is not a reasonable policy in instances in which it violates maximum cardinality.

A second reason for giving priority to (MC) over (ST) is that asylum is a public good and is as such in constant short supply (Fernández-Huertas Moraga and Rapoport (2014)). For example, remember that the total number of persons relocated in the EU up to July 2016 was less than 2% of the 160,000 people envisaged (European Commission (2016a)). On a larger scale, the UN Refugee Agency estimates the projected global resettlement needs in 2017 more than seven times higher (over 1.19 million) than the sum of the expected global quotas from resettlement countries (170,000, see UNHCR (2016)). The gap between demand and supply is so blatant that the desideratum to bring many refugees into a safe harbour trumps the desideratum to satisfy "as much as possible" the preferences of fewer.

4.2.3 When Preferences Can Be Taken into Account

It is also possible to investigate the conditions under which (ST) and (MC) are jointly satisfied. For example, Andersson and Ehlers (2018) design a mecha-

[12] E.g., Harsanyi (1976). Note that this is not a precise characterisation because the matching framework introduced here does not allow for cardinal utilities which are required for utilitarianism to be a meaningful doctrine. The labelling as "utilitarian" is a mere approximation.

nism that produces maximum stable matchings for assigning refugee families to landlords in Sweden. In that context, maximum stable matchings exists because refugees' and landlords' preferences only range over common languages spoken and the sizes of families, and they are assumed to be correlated in a specific way. In the present context of distributing refugees over countries, preference structures are more complex. However, a simple sufficient condition can be given: note that (ST) and (MC) conflict only if some countries deem some refugees unacceptable, or if some refugees deem some countries unacceptable. Recall that $E \subseteq R \times C$ denotes the set of acceptable refugee-country pairs, so that $E = R \times C$ indicates that everyone finds everyone acceptable. Then, we have the following simple fact.

Proposition 4.1 *In CA-instances in which $E = R \times C$, stable maximum matchings exist.*

It is easy to see that the proposition is true. Suppose $E = R \times C$ in a given CA-instance. It can then be shown that the stable mechanism μ^C provides a maximum matching. Applying μ^C, a refugee r_i accepts any country c_j's proposal in a step n unless in some step up to n, a preferred country has proposed to r_i. Recall that q_j denotes country j's quota. There are two cases to consider. First, suppose there are more places than refugees, i.e., $|R| < \sum_{k=1}^{m} q_k$. Each refugee who gets a proposal at some point will get matched because she only rejects if she already has a better offer. However, each refugee gets a proposal at some point because $E = R \times C$. Thus, the resulting matching M has a cardinality equal to the number of refugees, $|M| = |R|$, and so M is maximum. Second, suppose the number of refugees exceeds or equals the sum of all quotas, i.e., $|R| \geq \sum_{k=1}^{m} q_k$. Because $E = R \times C$ every country will fill its quota, and $|M| = \sum_{k=1}^{m} q_k$. Hence, M is again maximum.

Although the proposition states a very simple fact, it leads to interesting moral considerations. Can we require as a policy that every country deems

every refugee acceptable and vice versa? The former is the case in the context of the EU's relocation mechanism: member states are required to accept all types of refugees, not only those on their preference lists (Council of the European Union (2015)).[13] The reason this is required is that unilaterally declaring groups of refugees unacceptable may open the door to discriminatory policies. Moreover, it would allow countries to "play dummy": by stating a small or empty preference list, countries could participate in the system without fulfilling their quotas. It is reported in European Commission (2016a) that this has indeed happened in the context of relocations in the EU—although it is an illegal practice—and that member states are urged to refrain from it. Thus, for the initiators of the EU relocation mechanism, it seems uncontroversial that countries must accept all types of refugees, and the problem is rather to enforce this rule. However this may be enforced in practice, the CA model should be modified by imposing the restriction that the countries' preference profiles range over the whole set of refugees; formally, $A(c_j) = R$ for all $c_j \in C$.

To require that all refugees deem all countries acceptable is more problematic. As shown in the above example, if it is to be expected that in a given country the rights of a certain group of refugees are violated, then refugees who belong to this group can justifiably refuse to go there. Reasons for justified refusal include denial of non-discriminatory access to national services and public goods.

On the other hand, if it is the case that in each country within the system all refugees' rights are respected, refugees can be expected to deem all countries acceptable. Suppose the sum of the available places equals the number of refugees within the system. The following may then be considered a condition of fairness: a given refugee gets matched if she declares all countries accept-

[13] Cf. "Member States retain the right to refuse to relocate an applicant only where there are reasonable grounds for regarding him or her as a danger to their national security or public order", and: "Member States of relocation...should be ready to welcome all types of migrants (families, unaccompanied minors, single male applicants)" (Council of the European Union (2015)).

able.[14] In other words, if a refugee is willing to be matched with any country within the system she is guaranteed a place. This fairness condition agrees with the European Commission's proposal according to which refugees' successful applications to the relocation scheme do not imply a choice as to which country they move, but they do imply a relocation (European Commission (2016a)).

To sum up, if refugees' rights are respected in each country within the system, then it can be required that "everyone finds everyone acceptable", which guarantees that (ST) and (MC) can jointly be satisfied. However, whenever the set of countries includes a country where some minorities' rights are violated—although for other groups it may be a safe harbour—refugees that belong to these minorities cannot be expected to deem that country acceptable. (ST) should then be given up: in such instances, it is not a reasonable policy to take preferences into account. The trade-off between (ST) and (MC) can thus be interpreted as delimiting the area where preference satisfaction is a desirable policy goal. In the remainder of this article, it is assumed that the condition of no rights violations is met, and thus that a maximum stable matching exists.

4.3 Compliance with Higher-Order Policy Goals and Ethical Principles

In this section, it is argued that, even if the requirement of no rights violation is met and a modified CA model adopted in which "everyone finds everyone acceptable", this model must be rejected. The reason is that the CA model conflicts with the third desideratum encountered in Section 4.1: that the system should ban the possibility of "incorrect uses of preferences" (European

[14] The proof is a trivial extension of the one for proposition above: if there are refugees who deem some countries unacceptable, then a refugee who deems all countries acceptable gets matched to at least as good a country as in the case where all refugees deem all countries acceptable.

Commission (2016a)). An "incorrect use of preferences" is an expression of preferences which are in conflict with higher-order policy goals or ethical principles. Call this desideratum on the system *compliance* (COM). Compliance is also violated if agents can "game the system", i.e., achieve a more preferred matching by handing in preference lists strategically. It is a well-known result in matching theory that no stable mechanism is strategy-proof in the CA model (Roth (1982)). In the following, two additional lines will be discussed along which (COM) is violated in the CA model: the model enables the expression of impermissible preferences; and popular countries may be unduly favoured at the expense of less popular countries.

4.3.1 Impermissible Preferences of the Countries

So far, we have examined a model in which countries have strict and complete preferences over refugees. In practice, since the number of refugees may be large, countries cannot give strict preference lists over refugees. Rather, countries have preferences over groups of refugees that are identified through a classification according to properties the countries are interested in, such as profession, languages spoken, family status, urgency, etc. Suppose such a classification system is available. Countries are then indifferent between members of one and the same group, and a tiebreaker must be applied in order for the system to work.[15]

The task of designing a feasible classification system immediately gives rise to ethical problems: what properties of refugees can permissibly figure in countries' preferences? It has been argued that immigrants should generally not be selected on grounds of ethnicity, and that countries' preferences should

[15] This is usually a randomisation device, which has potentially problematic consequences. For some strict CA-instances obtained by breaking the ties the matchings produced by deferred acceptance algorithms may be Pareto dominated for the refugees by results of other possible tie breakings. The algorithms may be modified to solve for Pareto efficient matchings but only at the cost of strategy-proofness (Erdil and Ergin (2008); Abdulkadiroğlu et al. (2009)). This may pose a problem in practice; for simplicity, it will not figure in the arguments given here.

be restricted to range over "neutral" properties such as particular skills (Miller (2008)). This principle should arguably also be applied in the context of asylum. It agrees with the EU's relocation mechanism, which allows member states to express their preferences over refugees, albeit with "due respect of the principle of non-discrimination" (European Commission (2015)).

It is, however, hardly possible to entirely ban the use of ethically impermissible preferences in the CA model if there are countries in the system with such preferences. Restrictions can be imposed on the classification system—e.g., that preferences along ethnicity be forbidden. However, ethnicities also determine other features that are present in any reasonable classification system, such as mother tongue. Correlations between different such factors may make it possible to game the system, thereby violating (COM).[16] In line with the standard rationality assumption in market design that was discussed in Chapter 2, we assume that the market participants game the system whenever it is possible to do so.

Even if it were possible to design a classification system that cannot be manipulated in this way, countries' preferences may still conflict with ethical principles and higher-order policy goals. The EU relocation mechanism provides plenty of evidence. For example, European Commission (2016a) urges countries to express preferences in line with the policy goal "to facilitate integration of the relocated person in the Member State of relocation". However, "the majority of Member States use the preferences as a means to exclude possible candidates rather than to allow for a better matching process for better integration". As a consequence, it is demanded that "Member States of relocation should limit to the extent possible the preferences expressed" (ibid.). However, this stands in stark contrast to the rules of the relocation mechanism, which allow countries to express preferences in order to achieve

[16] Miller (2008) excludes this on grounds of "good faith". However, with regards to asylum, good faith is a weak hope to rely on in reality, particularly considering that many EU countries have not complied with their humanitarian responsibilities during the refugee crisis (Lücke (2016)).

good matches. It also stands in contrast to the CA model.

4.3.2 Unequal Treatment of Countries

Another important policy goal in the EU's relocation mechanism is to achieve equal treatment of countries. For example, concerning particularly vulnerable applicants, the Council urges "the necessity of ensuring a fair distribution of those applicants among Member States" (Council of the European Union (2015)). Equal treatment of countries is desirable both for fairness considerations and because it may be practically infeasible to impose a system in which some countries are worse off than others. If the system cannot ban the expression of preferences that leads to unequal treatment of countries, then there is a violation of (COM). The CA model may violate (COM) in this respect, in the following way.

In a usual asylum market, there are more and less popular countries for large proportions of refugees, that is, their preferences are correlated. If refugees' preference relations are sufficiently homogeneous, then deferred acceptance algorithms (including μ^C) in the CA model show favouritism to the most popular countries. For example, suppose there are two countries, HI and LO, and 100 refugees $r_1, \ldots, r_{100}$, all of whom prefer HI to LO. Applying μ^C, HI can "cherry-pick" its favourite group of refugees until its quota is satisfied. The problem is that LO's preferences are not at all taken into account: under any possible preference list, it will be assigned the same refugees. Moreover, this implies a practical problem: why should LO be willing to join the system? But if countries that consider themselves on the "losing side" are discouraged from participating in the system, this may produce a market that is too short on the supply side if participation is voluntary, or no market at all if they are co-policy setters with sufficient weight.

Are refugees' preferences really so homogeneous that popular countries would

be shown favouritism at the expense of unpopular countries? This question has not been answered conclusively. First of all, it is not clear how to set the threshold of when unacceptable favouritism begins. However, even if this is settled, so far not much data have been collected as to the actual preferences of refugees, and it is difficult to infer from the total or relative numbers of past asylum applications in a given country to the popularity of that country; for asylum seekers may not have had a choice as to where to apply for asylum. It seems a risky policy to implement a system in which popular countries are shown favouritism if refugees' preferences are sufficiently homogeneous, given that it is an open question whether their preferences are indeed so homogeneous.

Furthermore, there are indicators suggesting that refugees' preferences are relatively homogeneous. Two of the most important factors that shape refugees' preferences are family and diaspora in a country (e.g., Roth (2015c); on a local level also Katz et al. (2016)). These factors tend to cluster the preferences of a population of refugees from a given country or region and make them more homogeneous than those of populations from different countries or regions. But currently, the main population of asylum seekers is centred on few countries.[17] This may be evidence that their preference relations are indeed relatively homogeneous.

Jones and Teytelboym (2017a) argue that one can turn the tables on homogeneity of preferences. The risk of homogeneity may have positive effects on the market, so their argument goes, because countries will have incentives to court refugees in order to become popular destinations. This, however, seems to be an overly optimistic claim. For instance, a country may be unattractive for reasons that cannot be eliminated by changing its incentives—particularly, economic reasons. It seems that this country would then be unjustly disadvantaged in the CA model, if refugees' preferences are homogeneously biased

[17] In 2015, almost one out of three first time asylum seekers entering the EU originated from Syria, followed by Afghanistan (14%) and Iraq (11%) (Eurostat (2015, 2016)).

against it. Moreover, the argument does not address the problem that the market may not even come into existence if states are not interested in "courting" refugees. This will be the case especially if they consider themselves on the losing side as a consequence of past deterrence of refugees.

Moreover, the incentives argument can be turned around. As discussed in the previous section, refugees can be expected to deem a country acceptable only if their rights are respected there. However, a country interested in diminishing the number of refugees matched to it may have incentives to deter them from coming through drastic messages, or even to violate the rights of refugees already in the country in order to achieve this goal. Such policies are in effect in various countries all over the world,[18] and the fact that the CA model may enforce the incentives for such policies casts doubt on its normative desirability and effective operation.

Summing up, the CA model not only makes it difficult to prevent discriminatory policies. It also likely disadvantages some countries at the expense of others, thus provoking inequality among the countries and yielding undesirable incentive structures. The CA model violates (COM) in these respects, and, arguably, this is sufficient reason to reject the CA model for the asylum market.

4.4 Asylum as School Choice Problem

In the previous section, we encountered two lines along which the CA model violates (COM). First, even if the classification of refugees is restricted to "neutral" properties, it cannot be ruled out that some countries manipulate the system by expressing "incorrect preferences" (European Commission (2016a)). Second, popular countries' preferences likely receive overproportional weight

[18] For example, Hungary has been accused of criminalising and thereby violating refugees' rights (e.g. UNHCR (2015), or Amnesty International (2015)). Similar reproaches have been addressed to Australia (e.g. Fletcher (2014)).

at the expense of less popular countries.

What drives both problems is that countries are allowed to state preferences over groups of refugees. This motivates the following modification of the system. Refugees' preferences are taken into account, as before. However, countries are not considered economic agents with preferences over the people they provide with asylum; instead, asylum in a country is an object to be consumed by an asylum seeker. It is an object in short supply, but at the same time, it would be "repugnant" to sell it on a free market, in the sense of Roth (2007): many people think such transactions should not occur even if agents in the market would voluntarily engage in them.

How should asylum be distributed, if not through a free market? Arguably, a country should give its available places to the applicants who need it most or would most profit from it. This may be determined through *priorities* for specific *features of refugees* (instead of preferences over refugees). Priorities may comprise features such as vulnerability, urgency, dependants in a country, languages spoken, specific skills, etc. (these examples are policy goals set by the Council of the European Union (2015, 2016)). Criteria for setting priorities must be commonly agreed on, plausibly in conformance with a supranational institution such as the European Parliament, and should comprise only "neutral" properties. For example, a country could be allowed to prioritise refugees who speak its language but should not be allowed to prioritise race (Jones and Teytelboym (2017a)). National governments could then determine which features to prioritise while respecting the criteria agreed on.

In terms of matching theory, this suggests modelling asylum as a School Choice (SC) problem. The SC model was developed for the assignment of pupils to public schools in US school districts. In this context, schools are not assumed to be strategic agents, and it is only the pupils' welfare that matters (Abdulkadiroğlu and Sönmez (2003)). Formally, the SC model can be attained from the CA model by restricting the set of preference lists to the

refugees, and defining priority lists for the countries. Thus, an SC instance of a refugee-country matching problem is a five-tuple $(R, C, q, \mathbf{P}, \mathbf{Pri})$ with $\mathbf{P} = \{P(r_1), \ldots, P(r_n)\}$, and where $\mathbf{Pri} = \{Pri(c_1), \ldots, Pri(c_m)\}$ is the set of countries' priority lists. It is assumed that all refugees deem all countries acceptable and priorities range over all refugees, so $E = R \times C$. The definition of a matching is equivalent to that in the CA model.

Priority rankings are usually generated through a point system. If two applicants have identical points, the priority ranking may be determined through a lottery or continuous factors. In the context of the refugee match, it seems plausible to give a refugee who has been waiting longer for transfer more points in all countries' rankings than to a refugee with less waiting time spent, other things being equal; so *waiting time since registration in the system* could be used as a continuous variable to break non-strict priorities.[19]

Stability (ST) is in SC-instances usually interpreted as the elimination of justified envy (Abdulkadiroğlu and Sönmez (2003)), but, in the context of asylum, the term *elimination of justified grievance* may appear more appropriate. If a refugee does not get matched to her most preferred destination and the matching satisfies (ST), then she has a lower priority in that country than all the refugees matched to it and, hence, there is no ground for grievance. Deferred acceptance algorithms are applicable to SC-instances, and produce matchings that satisfy (ST). Let μ^R be the algorithm equivalent to μ^C but with the roles switched: the refugees propose and the countries accept the proposals in each step of the refugees with the highest priorities up to filling the countries' quotas. μ^R is typically preferred to μ^C in the SC model in which only refugees' welfare is taken into account because it produces the refugee-optimal stable matchings. Moreover, it is strategy-proof for refugees and thus strategy-proof

[19] Depending on the context, refugees could be registered in member states of the system or hotspots (reception centres in frontline states within the system), in camps external to the member states of the system, or even in diplomatic missions such as embassies in the region of origin (Ademmer et al. (2015)).

tout court in a model in which countries do not strategise.[20]

An immediate normative problem that arises when applying the SC model to the distribution of asylum is that in this model, stability does not imply efficiency (Abdulkadiroğlu and Sönmez (2003)).[21] The following example from Roth (1982) illustrates this.

Example 4.2 *There are three refugees,* r_1, r_2, r_3*, and three countries* c_1, c_2, c_3 *with* $q_{1,2,3} = 1$*. The refugees' preferences and priorities are specified in Table 4.2.*

Table 4.2: Priorities and preferences in example 4.2.

Pri	**P**
$r_1 \succ_{c_1} r_3 \succ_{c_1} r_2$	$c_2 \succ_{r_1} c_1 \succ_{r_1} c_3$
$r_2 \succ_{c_2} r_1 \succ_{c_2} r_3$	$c_1 \succ_{r_2} c_2 \succ_{r_2} c_3$
$r_2 \succ_{c_3} r_1 \succ_{c_3} r_3$	$c_1 \succ_{r_3} c_2 \succ_{r_3} c_3$

The unique stable matching is $((r_1, c_1), (r_2, c_2), (r_3, c_3))$*. It is Pareto-dominated by a matching where* r_1 *and* r_2 *switch their countries:* $((r_1, c_2), (r_2, c_1), (r_3, c_3))$*. This matching is not stable because* (r_3, c_1) *forms a blocking pair.*

Is asylum a context in which complete elimination of justified grievance should be ranked before efficiency, or vice versa? This depends on which interpretation of priorities is deemed appropriate in this context. Following Abdulkadiroğlu and Sönmez (2003), if the interpretation ought to be, "a refugee of higher priority in a country is entitled to asylum in that country before a refugee with lower priority", then we obtain elimination of justified grievance because there cannot be blocking pairs. If priorities are interpreted in a weaker sense and can be violated, exchanges as in Example 2 are possible which achieve efficient matchings but at the cost of producing blocking pairs.

[20] The importance of strategy-proofness in the context of locally matching refugees is discussed in Jones and Teytelboym (2018) and Delacrétaz et al. (2016 version).

[21] Note that this is the case only when we insist on strategy-proofness; there is no such tradeoff when countries strategise because stable outcomes are then efficient. See e.g., Kesten (2010). Thanks to an anonymous referee of the journal *Games* for raising this point.

While there may be some leeway for policy makers to decide which criterion to prioritise (cf. Jones and Teytelboym (2017a)), we anticipate that (ST) is important for the refugee match and normatively called for even if this leads to some efficiency loss. In the example, the efficient matching assigns r_3 to her least preferred country even though she has higher priority in her first-choice country, c_1, than r_2 who is assigned to it. Thus, the switch can be deemed unfair against r_3 because it causes r_3 to have justified grievance. (ST) blocks such unfair switches. This suggests that the elimination of justified grievance can be interpreted as a condition of fairness. There is a further argument to the effect that this fairness condition trumps efficiency in this context: we have a trimmed efficiency criterion that only takes into account the welfare of one side of the market. If switches are possible as in the example, then countries could well complain by asking, "why are priorities even collected in the first place", and refuse to accept matches that are a result of such switches. (ST), on the other hand, achieves fairness not only among refugees but also towards countries because their priorities for refugees are respected.

As mentioned in Section 4.2, in many contexts fairness also contributes to the thickness of markets, giving agents incentives to participate in the system. In many contexts, if (ST) is violated, then the lack of fairness leads agents to rematch and finally, to an unravelling of the system. In our context, agents could be forced to participate in the system, and rematchings could perhaps be made impossible (cf. Jones and Teytelboym (2017a)). Note also that of course fairness cannot prevent *illegal* secondary movements. Nevertheless, a lack of fairness would likely lead to frustration with the redistribution system, which may, in turn, reinforce incentives to engage in illegal movement to countries other than the one matched to, or to vanish from the system altogether after learning about the matched destination country. Moreover, if countries' priorities are public information, then it is relatively easy for refugees to check whether they are being treated unfairly. They would just need to ask others about their priorities and check whether they are matched to a more pre-

ferred country. Such information flows can be expected to be particularly high in the important stage when refugees are waiting in a hotspot or camp for their relocation, and when it is crucial that they have confidence in the fairness of the system. Finally, there may also be practical arguments that (ST) should be prioritised over efficiency. If justified grievance is not completely eliminated, then there may be legal appeals from individuals who have a justified grievance.

4.4.1 Discussion and Objections

The desiderata (ST), (MC), and (COM) can be jointly satisfied if priorities can be agreed to be imposed in a certain way. Since it is assumed that everyone finds everyone else acceptable, (ST) and (MC) can be mutually satisfied, and deferred acceptance algorithms solve for stable matchings of maximum cardinality (the argument in Section 4.2 extends to this case). Moreover, priorities can be formulated that satisfy (COM), and thus the two problems encountered in the previous section can be resolved in the SC model. First, criteria must be agreed on for priorities that are acceptable. For example, it seems reasonable that countries be allowed to prioritise refugees that speak their language (Jones and Teytelboym (2017a)). Some countries may still be keen to prioritise refugees on non-neutral properties such as religion; this is why it is crucial to clearly determine which types of priorities are acceptable and which are not. Doing so requires countries to give public reasons for the imposition of priority structures, and possibly mediation by a confederation-level institution. This would also prevent countries from being strategic, thereby adapting to the rules of the SC model.

Second, in addition to the specific needs of the refugees, relational factors between different host countries can be taken into account so as to avoid the danger of favouritism due to homogeneity. In the example where all refugees prefer country HI to country LO, suppose HI has a stronger economy than

LO, as measured in GDP per capita. Suppose, moreover, that all refugees are workers with identical skills but half of them, r_{51}–r_{100}, are war-affected and unable to work. For fairness reasons among the countries, humanitarian factors could be given a higher priority in HI and economic factors (such as integrability into the labour market) a higher priority in LO, thus matching overproportionally many refugees from r_{51}–r_{100} to HI and overproportionally many refugees from r_1–r_{50} to LO. This is again a process that calls for negotiation and mediation by a higher-level institution.

If priorities are imposed as described above, (MC), (ST), and (COM) can be jointly satisfied using deferred acceptance algorithms in the SC model. This model should thus be preferred to the CA model in the context of asylum. However, there is a counterargument: this comes at the price of treating asylum places as objects to be consumed by refugees. If a supranational institution is involved in the decision as to which types of priorities are acceptable and which are not, national policy makers may receive the impression that their national sovereignty is threatened under this model. On the face of it, it seems more realistic to model countries as agents with preferences because their governments clearly have preferences on issues of immigration. However, if the CA model appears more attractive to national governments than the SC model, it may be difficult in practice to impose the latter as an asylum policy whenever countries can voluntarily choose to participate in the system or have legislative co-determination.[22] This may, the objection goes, provide a justified reason for why the EU allows countries to express their preferences over refugees in the first place.

To counter this objection, remember what the European Commission seeks to achieve by allowing countries to express preferences over refugees: to "facilitate their integration into the Member State of relocation". However, be-

[22]This is the case in the EU: in the context of resettlements participation is voluntary and in the context of relocations participation is obligatory but countries have legislative co-determination through the Council.

cause of her private information, a refugee herself knows best where she will thrive; namely, "refugees go and integrate where they have family, where they have community, or where they think they can support themselves—in that order".[23] It can be assumed that this is also the member states' primary goal. However, for this objective, countries interested in their integration will share refugees' preferences to pool their families and communities and to integrate them into the labour market. The SC model achieves this by collecting refugees' preferences and allowing countries to prioritise certain groups of refugees along neutral properties.

A supranational mandate for the asylum system has additional advantages. Importantly, it permits a consistent definition of what counts as a justified asylum claim. A system that takes refugees' preferences into account is more attractive for refugees than a system that does not. Such a system can thus be expected to increase the demand for asylum. However, countries usually have an interest in narrowing down the market, rather than provoking a bigger run on asylum. An effective way to prevent this is to apply a clear-cut definition for the identification of justified asylum claims. However, this is difficult to agree on by single states with diverse standards for asylum. In a nutshell, we need a "communitarised" asylum system (Ademmer et al. (2015)). The SC but not the CA model takes a step in this direction.

4.5 Conclusions

Policy makers are increasingly interested in the question of *which* refugees to provide with asylum, in addition to the question of *how many*. The fact that mechanisms for distributing refugees can be designed in different ways gives rise to novel normative considerations. Reflecting on these has been the aim

[23] The CEO of the refugee resettlement agency *Hebrew Immigrant Aid Society* (HIAS), quoted in Roth (2015c). In a similar vein, Rapoport (2016) writes: "numerous studies show that the best indicator of future integration of migrants is the preference they express for a particular country".

of this chapter.

Some distribution mechanisms, such as the relocation mechanism in effect in the EU, seek to achieve good matches by taking countries' or refugees' preferences into account. I have argued that doing so may reduce the number of refugees matched if refugees may declare countries unacceptable or vice versa. However, to require that refugees accept any possible match presupposes that each refugee's rights be respected in all countries within the system—even in countries where they do not end up living.

Second, I argued that national governments should not express preferences over refugees. Instead, priorities for refugees should be imposed according to humanitarian factors and fairness considerations among the countries within the system. We then made the case that fairness is normatively called for and may in the context of distributing asylum beat efficiency considerations.

In terms of matching theory, I identified several reasons why asylum should not be modelled as a College Admissions but as a School Choice problem. Results from matching theory also helped to identify some of the problems that occur in the EU's relocation mechanism. For example, the fact that the mechanism does not systematically take refugees' preferences into account leads to unstable matchings and thereby violates fairness, which can be seen as a minimal condition for content. I hope to have shown that matching theory provides important tools for allocation problems in which agents' preferences matter and, thus, for numerous problems of distributive justice.

My aims have been normative: to reflect on the conditions under which it is a desirable policy to take preferences into account, and whose preferences to take into account. More technical work remains to be done for market designers, in particular to design fair mechanisms for matching refugees who migrate as families, and to investigate how relocation processes could be accelerated. Moreover, more work remains to be done for policy makers, to impose fairer asylum policies by drawing on insights from matching theory combined with

the kinds of normative considerations that have been discussed here.

Chapter 5

Kidney Exchange and the Ethics of Giving

Over 1.2 million people worldwide die of kidney disease each year, making it a little-noticed epidemic, which is comparable in scale to all deaths by road injuries.[1] Increasing the number of kidney transplants from live or deceased donors would save many lives, and it would improve the life expectancy and quality of many more. At present, there is a striking shortage of kidneys for transplantation.

Many patients have willing live donors, who cannot donate to their loved ones because they are biologically incompatible. Kidney exchange (KE) promises relief. For example, suppose your partner needs a kidney, but you cannot donate because you are incompatible. If the same is true of a different donor-recipient pair, it may be possible that you donate to the other recipient and your partner gets the other donor's kidney. Moreover, some people decide to donate *altruistically*, that is, they give a kidney to a stranger without receiving anything in return. Their gifts can trigger chains of KEs, thus multiplying the

[1] See Wang and et al. (2016), pp. 1483 and 1490 for kidney disease. They estimate the number of deaths by road injuries worldwide at 1.3-1.4 million annually (p. 1491).

benefits from a single donation. Below, a detailed description of different types of KE will be provided.

However, KE frequently meets ethical objections, in particular concerning donor protection. These are embodied in the transplant laws in many countries, which prohibit live organ donations to strangers, making the implementation of a broad range of KE procedures impossible. In light of the shortage of kidneys for transplantation, it is an urgent matter to clarify the ethics underlying KE.

This chapter aims to do this. It examines the implications for live kidney donations of some weak tenets from the ethics of giving. What I call the *effectiveness principle* is such a tenet. It says, roughly, that if a donation can be allocated to benefit some persons, or to benefit more persons than the first allocation, and the benefits to everyone are roughly identical, then we ought to choose the second allocation. This principle implies that, when an autonomous donor is given a choice between donating into a waiting list or into KE, in many cases morality requires the latter. KE is thus instrumental in meeting a moral obligation, which provides a novel argument for KE. We also consider the motivations for kidney donations. KE programmes maximise the goodness that a kidney donation can achieve; in particular, in their presence, a donation can trigger > 1 life savings. Therefore, KE may increase the motivation for donating, which constitutes a further argument for KE.

A balanced discussion must also take seriously the ethical concerns regarding KE programmes. Throughout, the German transplant law is used as a case study that expresses many of those concerns. I seek to distinguish the objections that are well-founded from those that are not and for the former, examine their implications for different forms of KE. It will be argued that even very conservative views on donor protection and distributive justice do not in principle oppose KE. Together, these arguments make a robust case in favour of providing a legal framework that allows implementing KE programmes.

This chapter is organised as follows. The next section argues for some weak principles from the ethics of giving. Section 5.2 explains the rationale for and basic procedures of KE. Section 5.3 examines the implications of the introduced principles for the allocation of kidneys, which constitutes the first argument for KE. Section 5.4 considers some of the critics' concerns and argues that they might restrict the domain and procedures of KE, but they do not in principle reject KE. Section 5.5 considers motivational aspects of live kidney donations, which constitute the second argument for KE. Section 5.6 concludes.

5.1 The Conditional Obligation to Donate Effectively

Before turning to the allocation of kidneys, some basic principles from the ethics of giving will be introduced. The philosophy of effective altruism has recently drawn increased attention to this ethical branch. Effective altruists believe that charitable giving should be done in a way that is most efficient in promoting the most good.[2] Effective altruism raises two questions: first, to what extent (if at all) is there a moral imperative for individuals to give resources to good causes; and second, if someone decides to give some of their resources to a good cause, what does morality imply for their allocation? We will only be concerned with the second, conditional question. To answer it, weak principles suffice that many people who are not effective altruists will find acceptable as well. The following trolley problem, which stems from Theron Pummer (see Pummer (2016)), may help to find these principles.

Example 5.1 (Trolley) *A trolley on a track A is headed for one innocent person, and another trolley on a track B is headed for 100 different, innocent*

[2]See Singer (2015) or MacAskill (2016). The former includes a chapter on altruistic kidney donations.

persons. Each trolley will kill everyone on its track with certainty - unless you stop it. You can stop one, but not both trolleys, by laying your arm on the respective track. If you do so, you will lose your arm, but everyone on that track will be rescued and nothing else will happen to you.

In line with the above, we make no claim as to whether there is a moral obligation to stop a trolley by sacrificing your arm. Instead, we are interested in the conditional question: if you choose to help, which trolley should you stop? Most will agree that you ought to stop trolley B. It would not be permissible to stop trolley A because you would only save one life instead of 100 by bearing the identical cost, namely your arm. Your rescuing would be terribly ineffective. This appears to be a robust moral intuition. For instance, suppose that, instead of sacrificing a limb, you can donate an amount of money to stop one of the trolleys (but not both). Or suppose that your donation (be it money, or your arm) will not save the lives of the persons on a track but would merely prevent them from having short, miserable, or diseased lives, where this benefit would roughly be the same for each beneficiary, no matter on which track. The felt conditional obligation to benefit many, rather than few, when the benefit is the same for each person, holds under a broad range of variations of the problem. Accordingly, we shall assume the following principle:

Definition 5.1 (Effectiveness Principle) *If one and the same donation can be allocated to benefit some persons, or to benefit more persons than the first allocation, and the benefit is roughly identical for each person under the two possible allocations, then, other things being equal, there is a conditional moral obligation to choose the second allocation.*

It is worth emphasising that the effectiveness principle, while implied by effective altruism, is weaker and therefore less controversial than the latter. For instance, effective altruism is demanding in terms of the cause that you should support. If fighting malaria brings about the most good in the most

cost-effective way, then effective altruism may imply that you ought not to support charities fighting homelessness in developed countries. The effectiveness principle would only make this recommendation if the benefits were roughly identical to each beneficiary and the beneficiaries under the former intervention more numerous. But typically, these benefits are unequal for the homeless in developed countries and persons suffering from malaria. Furthermore, effective altruism is typically thought to rely on a consequentialist and welfarist moral theory.[3] The effectiveness principle presupposes neither, as the conditional obligation implicated by it could instead be argued to arise, for instance, from fulfilling more claims to our aid.

The effectiveness principle is equipped with an other-things-being-equal clause in order to accommodate factors which can make a difference in various moral theories. For example, suppose the person on track A is a friend or family member. This and other "agent-relative" reasons break the symmetry between the two tracks, and they may excuse people from bringing about outcomes that are less-than-optimal from an impartial point of view.[4] There are also other possible circumstances that may block the consequent of the effectiveness principle, for example, if the person on track A would die with certainty whereas the persons on track B have a non-zero chance of survival, or if helping has bad side effects. The other-things-being-equal clause is supposed to capture all such relevant considerations.

I shall next extend the effectiveness principle minimally along three, independent lines. Each will be motivated by slightly altering the trolley problem.

[3] See Gabriel (2017). However, cf. Halstead et al. (2019).

[4] Horton (2017), p. 98. Absent agent-relative reasons, Horton agrees with Pummer that, if we are willing to make a sacrifice, we are morally obliged to bring about the best possible outcome by making this sacrifice. He gives a very similar version of the trolley problem, which differs in that stopping trolley B rescues all 101 persons. For example, this could be incorporated in the assumption that stopping trolley B also induces trolley A to stop, but not vice versa. This version is weaker because, if you deem it a conditional moral obligation to rescue the 100 persons on track B in Pummer's problem, it seems unjustifiable for you to hold that there is no conditional obligation to rescue everyone in Horton's problem. I choose the stronger version here because it is structurally similar to some cases of kidney donations we will be interested in.

(i.) We assumed above that your sacrificing an arm will stop a trolley with certainty. But suppose instead that your arm only slows a trolley down, stopping it later, so there is a small, positive probability that it will still reach any given person on its track. You have no reason to believe that this probability differs systematically for different persons, or different tracks: from your perspective, indistinguishable strangers stand at roughly the same distance on both tracks from indistinguishable trolleys that approach with roughly the same speed. Arguably, under this variation of the problem the effectiveness principle continues to hold. Since you cannot rationally differentiate between the probabilities of different persons being overrun or saved, the only morally relevant feature that distinguishes tracks A and B is the numbers of persons on them, just as before. So we shall assume that *the effectiveness principle holds in cases in which the benefits of the donation accrue to beneficiaries with probabilities smaller than 1, where there is no reason to believe that these probabilities differ systematically for the beneficiaries under the two allocations.*

It is worth noting that this extension of the effectiveness principle is weaker than a principle prescribing the maximisation of expected value. The latter would require sacrificing your arm on track B even when the probability of surviving for the persons on track B were much lower than for the person on track A. Our extension is silent about cases in which these probabilities differ.

(ii.) Suppose that your motivation for stopping a trolley is not the goal to rescue lives, but a different goal, such as meeting social expectations, or the desire to be seen as a hero. It seems that, even if such a non-altruistic motive is the driver behind the decision to help, it would still be morally wrong to rescue only one person instead of many. (Note that, in this case, you are not motivated by agent-relative reasons that would break the symmetry of the options.) So the thought experiment

advocates the effectiveness principle *even when the donor's motivation for donating is not altruistic and in the absence of agent-relative reasons that would break the symmetry of the options.*

(iii.) Finally, suppose a stranger is in the choice situation and you, instead of stopping the trolley yourself, observe her free choice. If she decides to stop a trolley, no matter on which track, you can prevent her from doing so. Suppose she decides to stop trolley B, thus rescuing 100 persons. Most of us will likely have the intuition that it would be morally wrong of you to stop her from doing so. (Perhaps the case would be more difficult should she decide to stop trolley A instead, but for our purposes this can be ignored.) There are limits to this intuition. For example, suppose the donor would not only lose her arm but would also bleed to death. Many think we should prevent people from sacrificing their lives. We shall assume that, *within reasonable limits, one ought not to prevent a donor from exercising her conditional obligation under the effectiveness principle.* We need not take up a stance here on how to set the reasonable limits in general, although we will touch on it below with respect to kidney donations.

For the rest of this chapter, the effectiveness principle and its three extensions will be assumed. We will employ them in the section after the next. Before doing so, we introduce kidney exchange in some depth.

5.2 Kidney Exchange and Altruistic Donations

Worldwide, there is a growing number of patients on waiting lists for kidneys. These are patients who suffer end-stage renal disease, that is, their existing kidneys fail. In most countries, there is a sizeable shortage of kidneys for transplantation. For example, in the US, 83,978 people were on the deceased

donor waiting list for kidneys in 2015,[5] 5,400 in the UK,[6] and around 8,000 in Germany.[7] The average time a person spends on these waiting lists is two and a half to three years in the UK,[8] almost four years in the US,[9] and around six years in Germany.[10]

In the meantime, many receive dialysis. But dialysis diminishes patients' quality of life and their life expectancy, and many die while on the waiting list. Moreover, dialysis is extremely expensive, thus putting a strain on healthcare services and it requires a medical infrastructure which is unavailable in many, especially developing countries (Wang and et al., 2016, p. 1525). Kidney transplants would extend many patients' life expectancy and life quality, and they are in most cases the cheaper alternative to dialysis. Thus, there is an urgent need to increase the supply of kidneys.

Healthy people have two kidneys and can donate one. Live donor kidney transplants offer the best prospects with respect to recipients' life expectancy and quality (e.g. Wallis et al. (2011)). However, because of incompatibility, which is mostly due to blood types, or specific antibodies of the donor, many willing donors are not eligible for donating to their loved ones.[11] KE promises relief for these patients. KE programmes seek to determine the matches between donors and recipients that maximise the number and quality of transplants. To ensure an informed discussion of the ethics of KE, I shall in the following

[5] `https://www.usrds.org/2017/view/v2_06.aspx`, last accessed on 04/11/2018. The number refers to dialysis patients only.

[6] `https://www.organdonation.nhs.uk/news-and-campaigns/news/nhs-blood-and-transplant-reveals-nearly-49-000-people-in-the-uk-have-had-to-wait-for-a-transplant-in-the-last-decade/`, last accessed on 04/11/2018.

[7] `http://statistics.eurotransplant.org/index.php?search_type=&search_organ=kidney&search_region=All+ET&search_period=by+year&search_characteristic=&search_text=`, last accessed on 04/11/2018.

[8] `https://www.nhs.uk/conditions/kidney-transplant/waiting-list/`, last accessed on 04/11/2018.

[9] `https://www.usrds.org/2017/view/v2_06.aspx`, last accessed on 04/11/2018.

[10] `https://www.dso.de/organspende-und-transplantation/warteliste-und-vermittlung/niere.html`, in German, last accessed on 04/11/2018.

[11] Biró et al. (forthcoming), referring to data from the Global Observatory on Donation and Transplantation, report that, "[d]epending on the country, 40% or more of recipients are incompatible with their intended donors" (p. 6).

introduce the basic procedures of KE in depth.[12]

The simplest procedure of KE is "two-way kidney paired donation", as shown in Figure 5.1, part (A). This procedure matches two incompatible donor-recipient pairs that are mutually compatible. So the donor of the first pair must be compatible with the recipient of the second pair, and the donor of the second pair compatible with the recipient of the first pair, in order for the exchange to happen. This requirement of *reciprocal compatibility* can be relaxed if additionally, more-than-two-way paired donations are feasible (Roth et al. (2007)). For example, part (B) in Figure 5.1 shows a three-way paired donation.

Other forms of KE alleviate the requirement of reciprocal compatibility by combining paired donations with *altruistic donations*. An altruistic, or "non-directed" donor is someone who gives a kidney to a stranger without receiving compensation.[13] In the absence of KE, altruistic donor kidneys are allocated to highly-ranked compatible patients on the waiting list. When combined with KE, an altruistic donor does not donate directly to the list. Instead, she donates to the recipient of an incompatible pair, whose incompatible donor simultaneously donates to the recipient of yet another incompatible pair, and so on, up to the last donor who donates to a patient on the waiting list. Since the altruistic donation kicks off various transplants, the resulting chains are called "domino chains" (Montgomery et al. (2006); Roth et al. (2006)). A domino chain is shown in Figure 5.1, part (C).

[12]The idea for kidney paired donation dates back to Rapaport (1986). The first exchange was realised in South Korea in 1991. Alvin E. Roth, Tayfun Sönmez, and M. Utku Ünver initiated a literature from a mechanism design perspective that searches for systematic procedures to increase the quantity and quality of transplantations (see Roth et al. (2004, 2005)). For a detailed history of KE, see Wallis et al. (2011).

[13] Altruistic donors do not typically specify the characteristics of the person whom they wish to receive their kidney, which is why "non-directed" and "altruistic" refer to the same class of donors. However, in some countries, e.g. the UK, directed altruistic donations are legal, that is, donors may donate to specific but unrelated persons, for example, a patient whose predicament was reported on TV. Directed altruistic donations may give rise to "repugnance" (Roth (2007)), for example if they produce markets for attention among patients with end-stage renal disease. However, they seem to be a marginal phenomenon and will not be considered here.

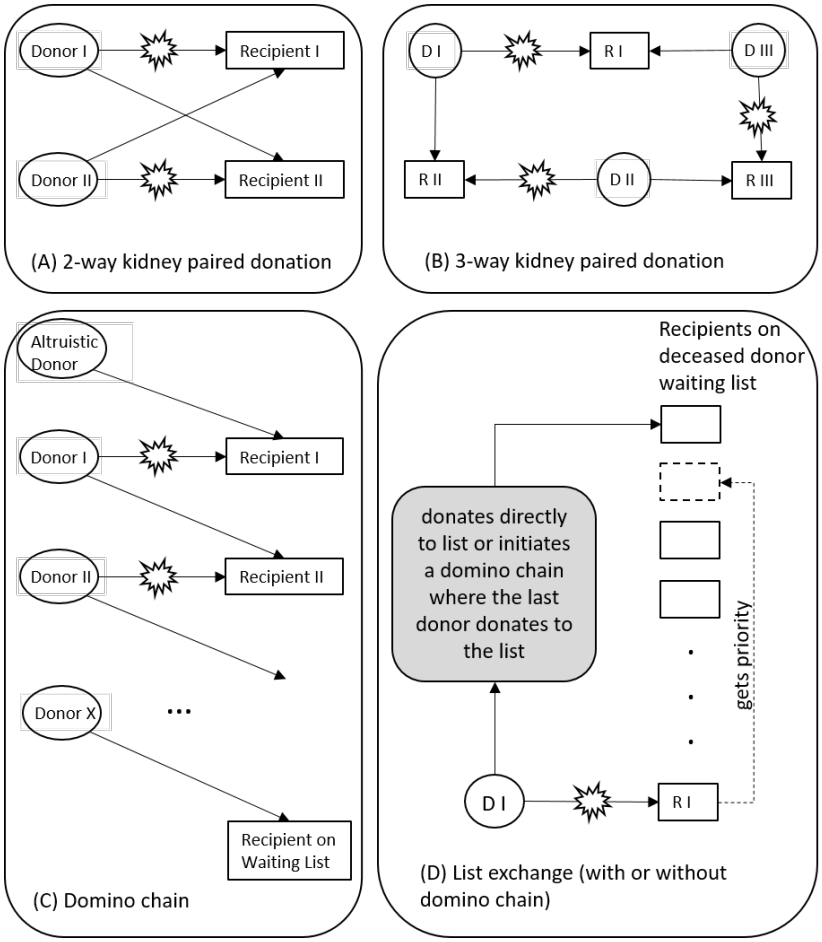

Figure 5.1: Simultaneous KE procedures. A solid arrow from A to B denotes an intended kidney donation from A to B. Exploding arrows denote incompatibility of the intended donor.

Similarly, paired donations could be combined with *list exchanges*, as in part (D) of Figure 5.1. In conventional list exchanges, the donor of an incompatible pair donates to a patient on the waiting list and in return her recipient gets priority on the list.[14] Instead of donating directly to the list, the donor could also donate to the recipient of another incompatible pair, thus kicking off a sequence of simultaneous exchanges, in which the last donor donates to the list and the recipient of the first donor gets priority on the list (Roth et al. (2006)).

Each of the procedures (A) through (D) triggers at least two transplantations, which are carried out *simultaneously*. This is because the promise to give a kidney is not legally enforceable, which poses the problem that in non-simultaneous chains, possible donors might renege on their promise to donate once their recipient received a kidney. The size of simultaneous chains is circumscribed, in particular by hospitals' logistics (each transplantation requires two operating rooms, and hospitals cannot accommodate many transplantations simultaneously), or by geography (kidneys must be transplanted as quickly as possible and should therefore not travel far).

However, there have been successful, "non-simultaneous, extended, altruistic donor" (NEAD) chains (Rees et al. (2009)). These consist in segments of domino chains, as shown in Figure 5.2. The last donor of a segment becomes a "bridge donor": instead of simultaneously donating to the waiting list, she initiates a new segment at a later date. There are two types of NEAD chains: closed NEAD chains specify a last donor, who donates to the waiting list simultaneously with the other donations of the last segment. Open-ended NEAD chains, in contrast, consist of indefinitely many segments. They end only when a bridge donor is ineligible to donate (e.g. because of a difficult-

[14] There are also list exchanges in which a donor donates now in exchange for a voucher that places her recipient on top of a waiting list in the future. This is viable when the donor's intended recipient does not need a kidney yet but can be expected to need one in the future, when the donor might be too old, or other mismatches might occur. For a news story about such a case, cf. `https://edition.cnn.com/2018/11/14/health/kidney-voucher-grandmother-donation-eprise/index.html`, accessed on 19/11/2018.

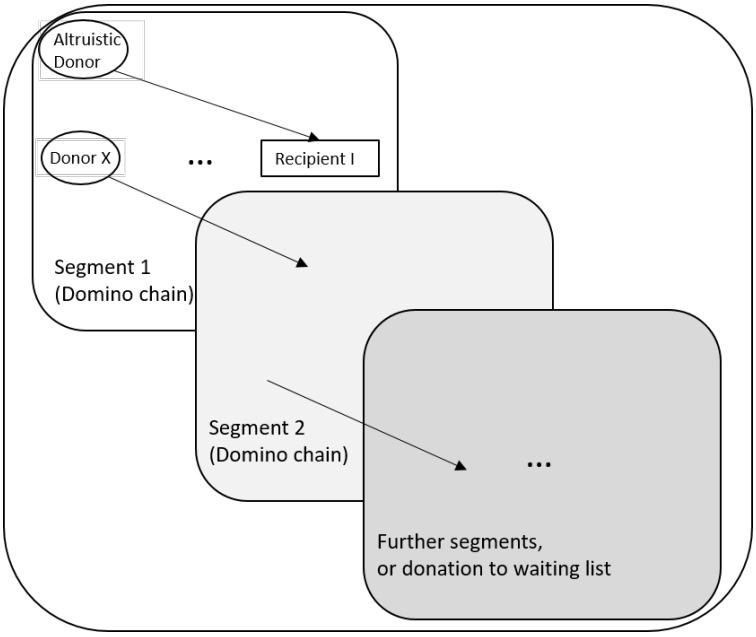

Figure 5.2: Non-simultaneous, extended, altruistic donor (NEAD) chain. An altruistic donor initiates a domino chain (Segment 1). The last donor (denoted X) from segment 1 becomes a bridge donor and initiates another domino chain (Segment 2) at a later date. The last donor of segment 2 either donates to the waiting list, in which case the NEAD chain ends; or she becomes a bridge donor and initiates segment 3, and so on.

to-match blood type), or if a bridge donor reneges on his promise to donate. Being non-simultaneous, NEAD chains alleviate the logistical obstacles that confine simultaneous chains. They promise a further increase in chain lengths (Ashlagi et al. (2011)), some of which have reached over 30 recipient-donor pairs in recent years. However, they raise ethical and motivational issues for the bridge donors, which we will encounter below.

KE programmes are being increasingly implemented in many countries (e.g. Roth (2015a, 2018), for Europe cf. Biró et al. (forthcoming)). Further developments that promise additional increase in the numbers and the quality of transplantations include global KEs, some of which have already taken place (e.g. Rees et al. (2017)); and the integration of compatible pairs into KE, which is possible if those pairs would profit from the exchange, or if they wish to engage in altruistic behaviour (Roth et al. (2008); Wallis et al. (2011)).

However, while KE is expanding, it is at the same time meeting ethical concerns. These are embodied in the transplant laws in various countries, which virtually ban KE programmes. For instance, German legislation permits donations only from persons of first and second degree of relationship, or who otherwise "manifestly stand in a special, personally close relationship" to the recipient.[15] This restriction makes most KEs and all altruistic donations, which are by definition provided by strangers, illegal. There have been piecemeal two-way kidney paired donations where the two incompatible pairs established personally close relationships with each other, thus allowing for legal paired donation. However, since the requirements on exhibiting such a relationship are high, and in the absence of an appropriate clearinghouse, success via this path is unlikely. This is even more so for three-way exchanges, the integration of which would achieve a more efficient use of the donor pool. Domino and list exchanges are rendered altogether impossible. The restrictive transplant law in Germany has led some patients to join KE programmes in other countries.[16] The law has met opposition, for example, in some recent newspaper articles,[17] but at the same time, various interest groups lobby in favour of retaining it.[18]

Germany is not an isolated case. In Europe, similarly restrictive laws are in

[15] Cf. paragraph 8 of the German Transplant Law, version from 04 September 2007 (BGBl. I S. 2206), changed by article 2 from 21 November 2016 (BGBl. I S. 2623).

[16] For example, for a report about a woman who went to Spain to donate a kidney to facilitate a transplant for her niece through KE, see `https://www.stern.de/gesundheit/organspende-niere-ringtausch-tausch-simone-reitmaier-6939720.html`. The patient's mother has since established a database for incompatible donor-recipient pairs to find compatible matches (`https://crossover-nierenspenderliste.de/`), and she is active in petitioning a change of the transplant law (`https://www.change.org/p/bundestag-gesetzes%C3%A4nderung-zur-einf%C3%BChrung-einer-datenbank-f%C3%BCr-die-%C3%BCberkreuzspende-von-nieren`). All webpages in German, and last accessed on 19/11/2018.

[17] E.g. `https://www.sueddeutsche.de/wirtschaft/forum-nierentausch-in-zeiten-des-mangels-1.2904824` and `https://www.taz.de/Debatte-Organspenden-in-Deutschland/!5519576/`, both in German, argue for an extension of the possible donor pool so as to allow broader KE. Both last accessed on 19/11/2018.

[18] For example, the interest group of live kidney donation, cf. `https://www.bundestag.de/blob/425002/51fac6cb911348a2c0723b971710919e/interessengemeinschaft-nierenlebendspende-e--v--data.pdf`, from 2016, in German. Last accessed on 09/01/2019.

effect in Bulgaria, Estonia, Finland and Hungary, among others.[19] Less but still relatively restrictive laws prevail, for example in Belgium, France, Greece, Poland and Switzerland, which are more permissive concerning kidney paired donation, but prohibit altruistic donations, thus ruling out domino chains.[20] Thus, there seem to be concerns in particular about altruistic donations that motivate many countries to impose restrictive transplant laws, which rule out various forms of KE.[21]

We shall uncover the possible objections to different forms of live kidney donations and their implications for KE in the section after the next. Before proceeding to that, we will examine KE in view of the effectiveness principle. For the time being, we shall assume that all types of live kidney donations, including altruistic donations to strangers, are feasible.

5.3 Kidney Exchange and the Effectiveness Principle

It won't come as a surprise what the effectiveness principle and its extensions imply for the allocation of live kidney donations. Yet we need to take care not to jumble different types of kidney donors. This section will consider the most important types of donors successively, and it will end by examining the implications for KE.

First, consider *altruistic donors*. When an altruistic donor donates to the waiting list, she may help one patient on that list. In the presence of KE programmes, altruistic donations trigger KE chains, thereby helping at least two, but possibly many more patients. From this and the effectiveness principle it follows that, when an altruistic donor is offered the choice between donat-

[19] Wissenschaftlicher Dienst (2017), p. 17 and Lopp (2013).

[20] Biró et al. (forthcoming), especially p. 12 and table 1.

[21] It is noteworthy that in the US, where altruistic donation is legal, some transplant centres have nevertheless been reluctant to accept these donors Tenenbaum (2016), p. 148.

ing into a waiting list or into KE, morality requires the latter.[22] Moreover, KE programmes use optimisation algorithms that maximise the number of possible matches within the pool of possible donors and recipients, subject to quality constraints (see below). The use of these algorithms guarantees that no possible allocation of kidneys in this pool could be more effective. Therefore, an altruistic donor donating into KE is thereby donating as effectively as possible.

So far, we have talked as if comparing successful transplants in the presence versus the absence of KE. But there is no guarantee for success: in a small number of cases graft loss, or other complications occur for the recipient. When an altruistic donor is offered the choice between donating into a waiting list or into KE, she does not know who would receive her kidney in each case and what their respective chances of success are. Before KE programmes were in effect, altruistic donations were typically allocated to highly-ranked patients on the waiting list in such a way that takes the match quality into account — including factors such as blood type compatibility, sensitization, age, and others. This increases the chances of success. Likewise, the optimisation algorithms used in KE programmes are programmed to maximise the quantity, but also the quality of matchings, where the latter may include all the factors that figure in the list allocations (Rees et al. (2009), p. 1100). So altruistic donors will have no reason to believe that the chances of success differ systematically for donating into KE versus donating to the waiting list. In this situation, the first extension of the effectiveness principle applies. According to this extension, the effectiveness principle is in effect when the benefits of the donation accrue with probability < 1 and there is no reason to

[22]Note that, as before, we do not make unconditional claims about whether, or in what situations, it might be morally obligatory to donate a kidney. It may never be morally obligatory to donate a kidney because the costs of doing so in terms of risks to health are substantial. Most theorists will likely agree that these costs exempt one from the obligation to donate a kidney. An exception might be Singer, who holds that altruism is obligatory to the point where the donors' sacrifices equal the recipients' benefits (e.g. Singer (1972)). This condition is typically satisfied in kidney donations, most clearly in cases in which donors don't experience complications or long term consequences, while the recipient would have died had they not received this kidney.

believe that this probability differs systematically for the beneficiaries under the two possible allocations. Thus, this principle requires donating into KE in the probabilistic case as well.

In contrast to altruistic donors, *directed donors* have agent-relative reasons that may block the obligation to donate into KE, as the other-things-equal clause of the effectiveness principle applies when a donor wishes to help a relative or friend. If they are compatible, it is usually uncontroversial that that person will receive the organ. If they are not compatible, it may nevertheless be possible for the recipient to receive a kidney through paired donation. So they may engage in KE and thereby achieve an effective allocation, but effectiveness results as a byproduct: it is not the effectiveness that requires the donor to engage in the paired donation, but the agent-relative reason that a donation to a stranger will provide a transplant for their loved one.

Next, consider *bridge donors in NEAD* chains. Remember that these donors donate to strangers after their recipient received their transplant. Similarly to altruistic donors, they either donate to a waiting list or they trigger a new segment within a NEAD chain. However, their motivation for donating differs from that of altruistic donors. A bridge donor donates to a stranger because she honours her promise to do so after her recipient received his transplant. She typically wouldn't have donated to a stranger otherwise, and would have donated to her recipient had he been compatible. This is why there are concerns that bridge donors may renege on their promise to donate once their recipients received their transplant, which will be discussed in the next section. Here, we note that the second extension of the effectiveness principle applies to bridge donors who deliver on their promise to donate. It stated that the effectiveness principle is indifferent to the motivation for the donation.[23] This implies that, no matter what their motivation, if a bridge donor

[23]The effectiveness principle ceases to apply if the motivation stems from agent-relative reasons. But bridge donors' reasons for donating are not agent-relative because their partners already received their transplants and they donate to strangers.

honours her promise to donate, and offered the choice to donate into a list or to trigger a chain, morality requires the latter.

It is less clear whether the same holds for *compatible pairs* when they are offered the choice to take part in KE. They may accept this offer for a range of reasons, for example that the recipient would profit from KE by receiving a better match than from her intended donor. Or they refuse the offer, for example when the compatible recipient prefers the organ of a related party to a stranger's. In these cases, their motivation is agent-relative and the effectiveness principle does not apply. In other cases, for example when they are indifferent, or partly altruistically motivated, the effectiveness principle does apply, requiring them to take part in KE.

Finally, consider the third extension of the effectiveness principle: it is not permitted to prevent a donor from exercising her conditional obligation to donate effectively. Concerning donations from the relevant groups above - altruistic donors, bridge donors, and some compatible pairs -, KE is instrumental in meeting their conditional obligation to donate effectively. Moreover, KE is the only way to achieve effectiveness.[24] It follows that it is morally wrong to prevent these donors from donating into KE. As a corollary, it is morally wrong to prohibit KE.

To sum up, weak principles from the ethics of giving have two important implications for live kidney donation: (i) there is a conditional obligation for altruistic and bridge donors, and for some compatible pairs, to donate into KE instead of into a waiting list, if they can choose to do so; and (ii) since KE is instrumental in meeting the conditional obligation to donate effectively, it is wrong to prohibit KE.

The effectiveness principle may have other implications that will not be con-

[24]This is certainly the case at present. We neglect here possible innovations in medicine that would overcome present immunological incompatibilities. We also neglect the possibility of monetary markets for organs, which would likely make altruistically motivated donations altogether unnecessary.

sidered in depth here. For example, it might be applied to make a case for global KE, which has the potential to substantially increase pool sizes and thus numbers and quality of transplants. However, global KE may give rise to separate ethical issues, for example, possible organ trafficking and unreliable medical care in developing countries, and it has generated opposition on these grounds (Delmonico and Ascher (2017)). These are difficult issues, worthy of a separate investigation, and will therefore not be considered here.

5.4 The Scope of KE and the Design of Transplant Laws

Where do the ethical concerns stem from, which are embodied in many transplant laws that virtually ban KE? There are various potential issues that were ignored in the argument from the effectiveness principle. We shall first discuss influential arguments against donations from strangers, which rule out most KEs. The Research Section of the German Federal Parliament provided a rich source of these arguments in a technical report.[25] We then discuss narrower arguments against specific forms of KE, especially NEAD chains. Finally, the implications of this discussion for the design of transplant laws will be considered.

5.4.1 Arguments Against Donations from Strangers

Protecting donors from possible harms. There is no evidence that live kidney donations significantly decrease donors' life expectancy, or quality of life (for a detailed discussion, cf. Tenenbaum (2016), p. 136 et seqq.). However, like any invasive surgery, they entail small risks of medical complications, including a very small, non-zero probability of death. These are possible

[25] Wissenschaftlicher Dienst (2017), especially p. 10 et seqq.

harms to healthy persons who receive no medical benefits from the surgery. Therefore, removing a kidney from such a person might be argued to violate physicians' duty to "do no harm". Accordingly, one of the reasons that the German transplant law prohibits donations from strangers is to protect live donors from such harms that their decision to donate might entail.

But risk does not imply harm. Prohibiting live donations on the basis of donor protection would require a duty to incur no risks of harm, or, more reasonably, no risks above certain thresholds, which must arguably be set relative to the benefits to the recipient. This is not the place to argue for a specific threshold that is acceptable for live kidney donations, but it may nevertheless be helpful to compare their risks to some other risks that many people face in their daily life. It is estimated that 3.1 per 10,000 kidney donors die during or within the first 90 days of their donation (Segev et al. (2010)). This mortality rate is comparable to working in refuse and recyclable material collection for a year, according to statistics on occupational hazards (Statistics (2017)). It is five times smaller than a year working in logging, which is listed as the most dangerous profession in these statistics. It has been argued, not least by medical practitioners, that these risks are reasonably low.[26] Moreover, they are arguably far outweighed by the benefits to the recipient. Furthermore, presupposing that donors are mentally healthy and not subject to coercion or exploitation (concerns that will be discussed below), it seems they have a right that their autonomous choice be respected (Cronin (2008)).

More central for our purposes is the fact that Germany and other countries allow directed donations while prohibiting donations from strangers. Their implicit assumption seems to be that the risk of harm to directed donors is justifiable but the risk of harm to anonymous donors is not. But these risks

[26]For example, Richard B. Freeman writes, "[w]e expose patients to all kinds of risks everyday for presumed benefits. Moreover, people willingly assume risks in their everyday lives, often much greater than those imposed by donor surgery, that have little or no direct benefit to their health. The risk that the harms from kidney donation will occur is very small compared with many risks we all face in everyday life" (Freeman (2012), p. 273).

do not systematically differ. So the claim that anonymous donors are more in need of protection from their decision to donate than directed donors must be based on other considerations than risk, which will be considered next.

Coercion. Germany also justifies the restriction of the possible donor pool with the need to rule out the possibility of coerciveness of donations and to secure their voluntariness. If it could be argued that anonymous donations entail an element of coercion that directed donations to family members and especially personally close persons do not, this would indeed constitute an argument for the restriction. To examine whether this is the case, we shall consider what a coerced donation could amount to.

In most countries, human kidneys are not for "valuable consideration". This is a legal term, meaning that it is prohibited both to donate and to receive kidneys in exchange for money, or other valuable goods or services. A promise is only legally enforceable if it is for valuable consideration. Thus, a promise to donate a kidney is not legally enforceable. This rules out the strongest form of coercion, which would subject the provision of a kidney to a legally binding contract. It also rules out the exploitation of the poor, as it is not possible to sell kidneys.

However, as legal scholars point out,

> "consideration is a slippery doctrine ... donors are allowed to direct that their kidneys be given to certain people: family members, friends, and others. This might seem like a transfer without valuable consideration, but that is not necessarily the case. The donor might transfer to such people rather than to a stranger because she expects to receive something in return—for example, household services or help in some other matter. Only a donation to an anonymous stranger could clearly be without consideration. Nonetheless,

> the common law of contract generally treats intrafamily transfers as occurring without consideration, and regulated entities and regulators have apparently taken this position for kidney donations to friends and family, as well." (Choi et al. (2014), p. 290 et seq.)

Thus, compensation and coercion are harder to rule out when someone donates to a close relative or friend. The organisation of KEs in countries where they are legal reflects this concern that personal relations can be instrumental for exercising coercion. Not only do transplant centres seek to rule out coercion through extensive background checks of potential donors, interviews, and education; it is typically also made impossible for mutually unacquainted persons in KEs to contact each other prior to the donation. There are various practical measures to enforce this, such as using different hospital sites. Some countries, e.g. Australia, discourage meeting even after the donation in order to rule out the possibility of posterior compensations, or of raising accusations, for example after graft loss.[27]

Alas, there is no guarantee that donations are always entirely free of some soft forms of coercion, in particular in emotionally close relationships. The argument that restricting the donor pool to especially close persons would help secure the voluntariness of a donation gets it the wrong way around. The altogether different conclusion here is that, if you want to allow directed, e.g. intrafamily donations - as most countries, including Germany, do - then there is no reason based on coerciveness for prohibiting anonymous donations, including altruistic donations.

There is a more subtle issue concerning coercion in NEAD chains, which will be considered below.

[27] Post donation, donors typically do have the right to know about the success of their donation. They may also learn about the characteristics of the recipients, and possibly the chain it triggered. This information could be, but is typically not given pre donation, because it might produce a feeling of coercion, for example when they realise that many people depend on their donation.

Slippery slope: KE and the commercialisation human organs. Germany also adduces the prevention of organ trade as a reason for the limitation of the donor pool. However, this argument is not convincing. It is empirically unfounded, as most countries condemn the practice of buying and selling organs and there is not a single country that has commercialised organ donations after implementing KE programmes.[28] Concerning black markets, there is no reason to believe that they are more likely to develop in the presence of KE. (It might be argued that the opposite is the case because KE helps to decrease the demand for kidneys.) Finally, the argument also commits the fallacy encountered before: why should it help for preventing valuable consideration in kidney donations to restrict the donor pool to especially close persons, where the risk of valuable consideration is higher?

5.4.2 Concerns about Specific Types of KE

Trade-offs between efficiency and fairness. As we have seen, without KE, altruistic donations are allocated to highly-ranked compatible patients on waiting lists. These lists incorporate medical, but also fairness principles, such as time already spent waiting, or priority of children over adults. Now, suppose that an altruistic donor decides to donate into KE instead of to the list. The resulting concern is most visible in open-ended NEAD chains. Remember that these are non-simultaneous chains, which end only when a bridge donor becomes ineligible, or reneges on his promise to donate. Thus, open-ended NEAD chains divert altruistic donations from the waiting list. Closed NEAD chains, in which a last donor is specified who will donate to the waiting list, are not necessarily subject to this diversion. But even in closed NEAD chains, the last donor might not donate to the list, for example if a bridge donor of an earlier segment reneged on his promise to donate.

When NEAD chains divert altruistic donor kidneys from the waiting list, they

[28] Iran is currently the only country where the sale of kidneys is legal.

might disadvantage those particularly vulnerable patients on the list who don't have living donors, because those patients are not eligible to participate in KE. On the other hand, these chains achieve large numbers of transplants. Thus, NEAD chains can be seen as promoting efficiency at the expense of fairness. The standard counterargument to this concern is that the efficiency that NEADs achieve helps all patients on the waiting list, namely by removing multiple patients from the list (Rees et al. (2009), p. 1100). Yet, diverting altruistic donations from the list may disadvantage at least some patients on the list. In particular, a patient does not profit if lower-ranked patients are removed from the list, so patients that are already highly-ranked can be expected to be disproportionately disadvantaged.

Unlike NEAD chains, other KEs do not in principle divert live donor kidneys from waiting lists. However, combining KE with waiting lists may disadvantage blood type O patients on the list. The reason is, very roughly, that blood type O patients can receive kidneys only from O donors, whereas O donors can donate to all blood types. Now, consider as an example a domino chain that an altruistic donor triggers and that ends with an incompatible donor donating to the list. The distribution of blood types among altruistic donors resembles that of the general population. Therefore, there is a high probability that this donor is O and she will donate to a hard-to-match O recipient. But it is unlikely that the incompatible donor who donates to the list is O, otherwise she would likely be compatible with her recipient. Thus, KE may systematically divert highly demanded O kidneys from the list.[29]

Some theorists suggest that the crucial ethical question concerning the trade-off between efficiency and fairness is this: how many additional transplants must the inclusion of altruistic donors into KE chains generate in order to justify the diversion of altruistic donors (in NEAD chains), or of blood type O altruistic donors (in general) from the waiting list? Transplant laws could do

[29] A similar concern arises in list exchanges (conventional or combined with domino chains). For a discussion, see den Hartogh (2010).

justice to a specific answer to this question by stipulating that the inclusion of altruistic donors into KE require a minimum number of transplants. Moreover, concerning the loss of O donors, they suggest "a requirement that, for every [altruistic donor] kidney donated to initiate a KE chain, a kidney of the same blood type must be donated to the [waiting list] at the end of the KE chain" (Woodle et al. (2010), p. 1464).

Risks for bridge donors. NEAD chains entail the risk that bridge donors renege on their promise to donate. However, the rates of reneging bridge donors appear to be small, and it has been argued that the utility benefits from NEAD chains outweigh these risks (Wallis et al. (2011); Tenenbaum (2016)).

There is yet another worry concerning bridge donors. NEAD chains are formed on the understanding that the bridge donors will donate to initiate a segment of transplants after their partners received their transplants (assuming that they continue to be medically and psychologically eligible and their circumstances have not changed substantially in the meantime). Because they gave this promise beforehand, they may feel obliged to donate after their recipients received their transplants. Bridge donors know that if they bail out, they thereby break the promise they gave, on the basis of which their partners received their transplant and on which various persons in need of kidneys rely. This might impose pressure on them, which may be felt as a form of coercion. It has been argued that it is morally problematic to put people in this position, and NEAD chains have been criticised on these grounds (see Tenenbaum (2016) for a discussion).

NEAD chains can also be argued to create coercion because the probability that matches are found is high. In general, potential donors can opt out at any moment before the transplantation without the transplant team revealing the reason. It is therefore possible for donors to bail out and falsely laying

the blame on medical reasons. This possibility is supposed to make the possibility of felt coercion less likely. However, NEAD chains make this excuse unconvincing, because in a NEAD chain the probability is much higher that a potential donor could go through with her donation (Wallis et al. (2011)).

On the other hand, it has been argued that NEAD chains can relieve donors from family pressure - since their recipient already received their kidney at the time of their donation –, which decreases the danger of coercion (Tenenbaum (2016), p. 156 et seqq.). Concerning the above worries, transplant centres select and educate possible bridge donors carefully. It may also be possible to relieve them from some of the felt pressure to be triggering a great number of transplants, simply by not telling them how long the chain will be prior to their donation. Moreover, since evidence suggests that the amount of felt coercion increases with time, time limits can be set within which their donation should happen, otherwise their promise is void. A more conservative solution would be to restrict KE to the simultaneous cases.

To sum up, we found the principled arguments against donations from strangers, which preclude most types of KE, wanting. However, it might be ethically required to restrict the scope of KE procedures. In general, the efficiency gains from allowing broader KEs must be weighed against increasing concerns with respect to the diversion of altruistic, especially type O donors from waiting lists and, in the case of NEAD chains, the potential felt coercion of bridge donors. My aim was not to argue for a specific weighting. Instead, the argument is the following. Suppose we take a conservative view and put heavy weight on avoiding (O) donor loss and bridge donors' felt coercion. A transplant law embodying this view might restrict or prohibit NEAD chains. It may also require that KE chains divert altruistic donations from the list only when they achieve a large number of transplants, and it may prescribe the prevention of O donor loss. The result would be a transplant law that places heavy weight on

donor protection and allocative fairness with respect to patients without living donors. The point is that this legislation would not even resemble the transplant laws we encountered earlier, such as the German transplant law, which require a personally close relationship between donor and recipient. Even a conservative view on live kidney donation, if sound, does not in principle reject KE programmes.

5.5 The Attraction of Effectiveness

KE may generate motivational benefits for donors. Consider the following report from Dylan Matthews, who altruistically donated his kidney in 2016:

> "...the very same day that I donated, [the recipient's] relative had their kidney taken out as well and flown to the West Coast. This second recipient also had a friend or relative agreeing to an exchange; so did the third recipient, who got the second recipient's friend's kidney. Our chain will let people enjoy 36 to 40 years of life they would've otherwise been denied.
>
> Our four kidneys were pretty good, but some chains can go even longer. A chain started by a 44-year-old man in California named Rick Ruzzamenti wound up getting 30 people kidneys. Ruzzamenti's chain let people live 270 to 300 years longer. You can literally measure the years of life his kidney donation chain gave in centuries."[30]

Matthews does not go so far as to suggest that he, or Ruzzamenti, decided to donate because of the potentially large numbers of life years that their donations would enable. But the passage provides clear evidence for the awe

[30] From `https://www.vox.com/science-and-health/2017/4/11/12716978/kidney-donation-dylan-matthews`, 2017, last accessed on 02/11/2018.

that donors experience when considering the large impact of their donations in terms of life years gained. This naturally suggests the hypothesis that, other things being equal, a donor's motivation is higher if the possible number of transplants triggered, or of life years saved, is higher.

This hypothesis, if true, has implications for KE. As we have seen, KE programmes use optimisation algorithms that maximise the number and quality of transplants. Thereby, they increase the number of lives saved, or of life years gained.[31] It follows that, if the motivation for donating is partly determined by and increases with the impact of the donation, KE increases the motivation for donating. The argument would apply to all donors who are partly motivated by altruism. This includes altruistic donors, but also other types - such as directed donors, bridge donors, compatible pairs -, as they may often be partly motivated by altruism as well.

In light of the striking shortage of kidneys for transplantation, if KE programmes promote the emergence of altruism, this constitutes a significant advantage. Thus, the hypothesis that the motivation for altruistic kidney donations is partly determined by the amount of good that they can be expected to achieve, if true, constitutes a second, motivational argument for the implementation of KE programmes.

This is not the place to investigate whether the hypothesis is true. It is an empirical hypothesis that could be confirmed by comparing trajectories of altruistic donations in countries where KE programmes exist to countries where they don't. We note here merely that the available evidence is consistent with the hypothesis. In many countries in which centralised KE programmes exist, e.g. in the US and the UK, the numbers of altruistic donations have been increasing in recent years.[32] More generally, there is evidence that donating

[31] Different algorithms may have unequal implications concerning numbers of lives saved and of life years gained. We can neglect this point here because an altruistic donation will typically increase both variables if KE is in effect as compared to the default of no KE.

[32] For data on altruistic donations in the UK, see Robb et al. (2018). For the US, e.g. Tenenbaum (2016).

effectively can boost donors' motivation.[33] We conclude that it is a reasonable hope that KE promotes the emergence of altruism.

In contrast, transplant laws that restrict the donor pool to relatives and persons manifestly close to recipients may involve problematic incentives. They risk conveying the image that there is something unethical about the gift to a stranger. Moreover, people hoping to go through with a kidney paired donation will have incentives to pretend that there are personally close relationships even when in reality there aren't. But such incentives to "game the system" cannot be in the interest of legislative authorities, and they are detrimental for building trust in the system. These motivational considerations speak against prohibiting live organ donations to strangers, and in favour of combining them with KE.

5.6 Conclusion

Weak principles from the ethics of giving make a strong case for KE programmes. These programmes are instrumental in allowing kidney donors to meet their conditional moral obligations that are implied by those principles. Therefore we ought not to preclude people from fulfilling these obligations by banning KE. There are possible ethical reasons for restricting specific procedures of KE, but they do not in principle reject it. Finally, KE may achieve motivational benefits that constitute a further argument in its favour.

The arguments given here are not wedded to a specific moral theory. They will appeal to effective altruists, but because of their weak, conditional premises, many people who are not committed effective altruists will welcome them as

[33] Parbhoo et al. (2018), p. 21. In a survey, 85% of donors revealed they paid very close attention to effectiveness when giving to charities (`http://static1.squarespace.com/static/55723b6be4b05ed81f077108/t/566efb6cc647ad2b441e2c55/1450113900596/Money+for+Good+I.pdf`, accessed on 10/12/2018). They also find that, even though they care about effectiveness, few donors spend time investigating the effectiveness of the charities they give to. So admittedly, the evidence is somewhat mixed.

well. They are also consistent with conservative views on donor protection and allocative justice concerning patients on waiting lists. I hope that these arguments will lead to a clarification of the debates on the ethics underlying KE, particularly in countries that have hitherto banned it.

This study calls for various follow-up projects. First, we explicitly excluded global KE, which has the potential to substantially increase the numbers and the quality of transplants. Ethicists are called for to weigh these benefit against the concerns that have been raised about global KE, for example, whether the risk of organ trafficking can be ruled out sufficiently in some developing countries. Second, the hypothesis about donors' motivation on which our argument from attractiveness drew should be investigated empirically. Third, in many countries that currently prohibit live donations to strangers, changes to legislation, for which we argued here, may turn out to be infeasible in the short term. In the meantime, work remains to be done concerning the implementation of some "slim" forms of KE programmes in those countries. For instance, restrictions of the donor pool to persons that are emotionally close to the recipient provide the possibility to match donor-recipient pairs that could in the next step meet in person and establish the required relationship. This would enable some forms of KEs, in particular kidney paired donations, which are currently conducted only sparsely in those countries. Making the most of existing transplant laws would improve the predicament of many people suffering from kidney disease, but it does not excuse decision makers' inaction.

Chapter 6

Conclusion

The first part of this thesis examined the foundations of economic design and I argued that they provide insights into important philosophical debates. In chapter 2, these were debates about the nature of rationality and of institutions. I showed that institutional designers' core constraint of incentive compatibility sits uneasily with a family of unorthodox theories of rationality in games that have been proposed. I argued that this leaves proponents of those theories with three options, each of which looks rather unappealing: give up the constraint of incentive compatibility, treat people as irrational, or give up their own standards of rationality in the context of institutional design. I also suggested an amendment to a recent account of institutions, namely that it should include reference to mechanisms. Furthermore, the standard meaning of "institution" allows for incentive-*in*compatible institutions.

In chapter 3, I analysed in depth the reform of a matching market for medical residents, which hasn't been considered in the philosophy of science to date. In cases like this, economic engineers seek to generate counterfactual knowledge about how creating or changing a market would bring about or alter its associated market outcomes. Using causal graphs, I gave an account of how this counterfactual knowledge is generated. According to my account, models,

e.g. from game theory, allow formulating policy goals that these outcomes should implement. Complemented with empirical methods, these models are used to devise institutions that bring about those goals. I also argued that the creation of knowledge in economic engineering shows that it may not be possible to evaluate from the philosopher's vantage point how economists should use models in economic design, and doing so might even do harm. A recent proposal in the philosophy of economics to the contrary is presumptuous and should be resisted.

In the second part of the thesis, we considered ethical aspects of specific matching markets. In chapter 4, I argued that basic results from matching theory provide insights into the distribution of refugees over host countries. There may be trade-offs between satisfying refugees' or countries' preferences and the numbers of refugees assigned, as well as between the efficiency and the fairness of matchings. Implementing distribution mechanisms requires careful weighing of these criteria.

In contrast to the specific distribution mechanisms that we examined for asylum, the problem we studied with regard to kidney exchange was more coarse-grained, because many countries ban this practice. In chapter 5, I tried to show that this status quo should be changed. I gave two novel arguments for the implementation of kidney exchange programmes: first, these programmes are instrumental in meeting a moral obligation, namely to donate effectively; and second, they may increase the motivation for altruistic donations, because the donation of one kidney may trigger > 1 life savings. We also examined ethical concerns that are embodied in transplant laws preventing the implementation of kidney exchange, and I argued that they can be overcome.

There are many questions with regard to these ethically loaded matching problems that are yet to be answered. Concerning the distribution of refugees, some recent studies have proposed matching them to localities through data-driven algorithmic assignments. Bansak et al. (2018) argue that using such

algorithms could significantly improve the integration of refugees into labour markets. The ethical implications of these proposals should be examined. In particular, it may be ethically problematic to distribute refugees algorithmically without taking their preferences into account. Concerning the allocation of live donor kidneys, an important question is whether or not global kidney exchanges that include donor-recipient pairs from developing countries are ethically justifiable. Moreover, should kidney exchange programmes eventually be replaced with a monetary market?

Of course, there is a multitude of problems for economic design in contexts other than the allocation of asylum or of organs. For instance, climate change presents problems which economic design can contribute to overcome. The design of climate treaty negotiations and cap-and-trade systems are but two examples. It is vital to keep in mind the lessons learned in chapters 2 and 3 when designing these systems, but unfortunately, these lessons have often been neglected. For example, the rules of many climate treaty negotiations are not incentive-compatible. Often, the global carbon emission abatement plans are simply the result of independent pledges of individual countries. But this does not align countries' incentives with the social goal to transition to a net-zero carbon economy in due time, but provides opportunities to free-ride instead. An incentive-compatible mechanism would condition individual countries' pledges on the pledges of the other countries. For an illustration, suppose all countries agree that collectively cutting down emissions would be better for everyone than business as usual, while individual defection would be better for the defector than cutting down emissions. Then the following mechanism is incentive-compatible: a mediator asks all countries to make pledges and commits every country to act on the least ambitious pledge of all countries (see MacKay et al. (2017), chapter 2).

While these are important domains to which economic designers might increasingly contribute in the future, I shall conclude with a few remarks on the

possible return of large-scale design. There are some signs of this happening. Posner and Weyl (2018) project a radical transformation of society through the application of mechanism design. They argue that private property is inherently monopolistic and should be abolished. They propose the common ownership self-assessed tax (COST) to replace it. This proposal, inspired by the ideas of Henry George, involves the utilisation of auctions on a very large scale. In this system, people and corporations assess the value of all major commodities that they wish to hold. The highest bidder wins the right to use the commodity. Users of commodities pay a relatively high tax on them (in the neighbourhood of 7 percent annually), which can be used to fund public goods and a basic income for every citizen. The right to use commodities is constantly auctioned: if someone else values a good more highly, she can acquire it, henceforth paying the same tax rate on her valuation. This system provides agents with incentives to reveal their true valuations and it achieves the efficient allocation of commodities. The authors argue that the advantages of this arrangement would be substantial, among them: the elimination of market power by turning markets in private property into markets in uses; increased public revenues; increased equality; increased innovation and investment in large scale projects; the transformation of private wealth into social wealth, which might make people less materialistic.[1]

Their proposal is interesting not only from the perspective of political economy, but also methodologically. I argued in chapter 1 that we can distinguish

[1] Posner and Weyl also propose a transformation of democracy through quadratic voting. In this system, important social decisions, e.g. about the provision of public goods, are made through referenda. Every citizen receives an equal amount of "voice credits" annually, which they can use on referenda in that year, or can stockpile to use in future referenda. They can "buy" any number of votes with their voice credits, but the costs are calculated according to a quadratic formula: 1 credit buys 1 vote, 4 credits buy 2 votes, and so on. Unlike current voting systems, QV takes into account the intensity of people's preferences – you can spend as many credits as you own on referenda that are important to you. Therefore, people will exercise influence in realms which they care about and in many cases know more about than the average citizen. Since under quadratic voting rational agents will spend their credits in proportion to how important the various referenda are to them, the system has the potential to make the provision of public goods efficient, analogously to what free markets achieve for private goods under ideal conditions.

two grand traditions of economic design: the first seeks to provide a framework for comparing and evaluating economic systems, but these comparisons are sometimes ambiguous and they remained mainly theoretical. The second is focused on the details of small-scale problems. It approaches design with an engineering mindset and is closer to policy-making, but this practical advantage came at the expense of losing the role of evaluating fundamental economic institutions. Posner and Weyl refocus on large-scale reforms. However, they also take into account lessons from economic engineering. Accordingly, their proposals do not simply go back to early design economics, but lift it to a new level. For example, the COST system, while constituting a radical reform of our economic system, could initially be applied for items that governments auction to the corporate sector, such as the radio spectrum. This connects to existing work in auction design, a field to which Weyl has contributed and which has been a successful area of application of economic engineering. So this implementation strategy would accommodate insights from economic engineering.

As economic engineers are gaining a better understanding of how real-world institutions work, they become better equipped for larger-scale social reforms through economic design. This is because understanding the details of specific interactions appears to be a critical precondition for successful large-scale design. For instance, marketplaces in a market economy are embedded in larger markets, which are in turn embedded in the economic system. Changing the fundamental institutions will also change markets and marketplaces, for example, by affecting the choices available to agents interacting in those markets. The growing body of knowledge about particular marketplaces and markets may improve the chances for successful large-scale design. The recent refocus on the large scale might be a sign that the discipline has matured to a point at which reformers can rely on system design. Whether this is the case remains an open question at present.

Bibliography

Atila Abdulkadiroğlu and Tayfun Sönmez. School Choice: A Mechanism Design Approach. *The American Economic Review*, 93(3):729–747, 2003.

Atila Abdulkadiroğlu, Parag A. Pathak, and Alvin E. Roth. Strategy-proofness versus Efficiency in Matching with Indifferences: Redesigning the NYC High School Match. *American Economic Review*, 99(5):1954–1978, 2009.

Esther Ademmer, Toman Barsbai, Matthias Lücke, and Tobias Stöhr. 30 Years of Schengen: Internal blessing, external curse? *Kiel Policy Brief*, 88, June 2015.

Anna Alexandrova. Connecting Economic Models to the Real World: Game Theory and the FCC Spectrum Auctions. *Philosophy of the Social Sciences*, 36(2):173–192, 2006.

Anna Alexandrova. Making Models Count. *Philosophy of Science*, 75:383–404, 2008.

Anna Alexandrova and Robert Northcott. Progress in Economics: Lessons from the Spectrum Auctions. In Harold Kincaid and Don Ross, editors, *The Oxford Handbook of Philosophy of Economics*. Oxford University Press, Oxford, 2009.

Anna Alexandrova and Robert Northcott. It's just a feeling: why economic models do not explain. *Journal of Economic Methodology*, 20(3):262–267, 2013.

Anna Alexandrova and Robert Northcott. Prisoner's Dilemma doesn't explain much. In Martin Peterson, editor, *The Prisoner's Dilemma*, pages 64–84. Cambridge University Press, Cambridge, 2015.

Amnesty International. Urgent Action: Hungary violates human rights of refugees, 2015.

Tommy Andersson and Lars Ehlers. Assigning Refugees to Landlords in Sweden: Efficient Stable Maximum Matchings. Working Paper, Department of Economics, School of Economics and Management, Lund University, 2018.

Erik Angner. Did Hayek Commit the Naturalistic Fallacy? *Journal of the History of Economic Thought*, 26(3):349–361, 2004.

Erik Angner. *Hayek and Natural Law*. Routledge, Abingdon, Oxon, and New York, NY, 2007.

Itai Ashlagi, Duncan S. Gilchrist, Alvin E. Roth, and Michael A. Rees. Nonsimultaneous Chains and Dominos in Kidney-Paired Donation – Revisited. *American Journal of Transplantation*, 11(5):984–94, 2011.

Robert Aumann and Adam Brandenburger. Epistemic Conditions for Nash Equilibrium. *Econometrica*, 63(5):1161–80, 1995.

Robert J. Aumann. What Is Game Theory Trying to Accomplish? In Kenneth J. Arrow and S. Honkapohja, editors, *Frontiers of Economics*. Basil Blackwell, Oxford, 1985.

Michael Bacharach. *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton University Press, Princeton and Oxford, 2006. Edited by Natalie Gold and Robert Sugden.

Kirk Bansak, Jens Hainmueller, and Dominik Hangartner. How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. *Science*, 354(6309):217–222, 2016.

Kirk Bansak, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner Duncan Lawrence, and Jeremy Weinstein. Improving refugee integration through data-driven algorithmic assignment. *Science*, 359(6373): 325–329, 2018.

Anna Bartsch and Luc Bovens. Towards a fairer distribution of asylum seekers. *voxeurop*, 2016. URL `http://www.voxeurop.eu/en/content/article/5041680-towards-fairer-distribution-asylum-seekers`.

Timothy Besley. What's the Good of the Market? An Essay on Michael Sandel's What Money Can't Buy. *Journal of Economic Literature*, 51(2): 478–495, 2013.

Ken Binmore. Institutions, rules and equilibria: a commentary. *Journal of Institutional Economics*, 11(3):493–496, 2015.

Kenneth Binmore. *Playing for Real: A Text on Game Theory*. Oxford University Press, Oxford, 2007.

Péter Biró, Bernadette Haase-Kromwijk, Tommy Andersson, Eyjólfur Ingi Ásgeirsson, Tatiana Baltesová, Ioannis Boletis, ..., and Joris Klundert. Building Kidney Exchange Programmes in Europe - An Overview of Exchange Practice and Activities. *Transplantation*, ahead of print, forthcoming.

Luc Bovens. The Ethics of Nudge. In Till Grüne-Yanoff and Sven Ove Hansson, editors, *Preference Change: Approaches from philosophy, economics and psychology*. Springer Science & Business Media, Berlin and New York, 2009.

Luigino Bruni and Robert Sugden. Reclaiming virtue ethics for economics. *Journal of Economic Perspectives*, 27(4):141–164, 2013.

Bruce J. Caldwell. The Emergence of Hayek's Ideas on Cultural Evolution. *Review of Austrian Economics*, 13(1):5–22, 2000.

Nancy Cartwright. If No Capacities Then No Credible Worlds. But Can Models Reveal Capacities? *Erkenntnis*, 70(1):45–58, 2009.

Stephen J. Choi, Mitu Gulati, and Eric A. Posner. Altruism Exchanges and the Kidney Shortage. *Law and Contemp Probs*, 77(3):289–296, 2014.

Peter Coles, John Cawley, Phillip B. Levine, Muriel Niederle, Alvin E. Roth, and John J. Siegfried. The Job Market for New Economists: A Market Design Perspective. *Journal of Economic Perspectives*, 24(4):187–206, 2010.

Council of the European Union. Council decision 2015/1523 establishing provisional measures in the area of international protection for the benefit of Italy and Greece. 2015. URL `http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015D1523&from=EN`.

Council of the European Union. Council decision 2015/1601 establishing provisional measures in the area of international protection for the benefit of Italy and Greece. 2016. URL `http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/european-agenda-migration/proposal-implementation-package/docs/20160321/provisional_measures_area_international_protection_benefit_italy_and_greece.pdf`.

Court of Justice of the European Union. Judgment in Joined Cases C-643/15 and C-647/15 Slovakia and Hungary v Council. *Press Release No 91/17*, 2017. URL `https://curia.europa.eu/jcms/upload/docs/application/pdf/2017-09/cp170091en.pdf`.

Antonia J. Cronin. Allowing autonomous agents freedom. *Journal of Medical Ethics*, 34(3):129–132, 2008.

Partha Dasgupta, Peter Hammond, and Eric Maskin. The Implementation of Social Choice Rules: Some General Results on Incentive Compatibility. *Review of Economic Studies*, 46:185–216, 1979.

David Delacrétaz, Scott Duke Kominers, and Alexander Teytelboym. Refugee resettlement. *Unpublished*, 2016 version. URL `http://www.t8el.com/jmp.pdf`.

Francis L. Delmonico and Nancy L. Ascher. Opposition to irresponsible global kidney exchange. *American Journal of Transplantation*, 17:2745–2746, 2017.

Govert den Hartogh. Trading with the Waiting-List: the Justice of Living Donor List Exchange. *bioethics*, 24(4):190–198, 2010.

Esther Duflo. The Economist as Plumber. *American Economic Review*, 107 (5):1–26, 2017.

Aytek Erdil and Haluk Ergin. What's the Matter with Tie-Breaking? Improving Efficiency in School Choice. *American Economic Review*, 98(3):669–689, 2008.

European Commission. Proposal for a Regulation of the European Parliament and of the Council. *COM(2015) 450 final*, 2015. URL `http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/european-agenda-migration/proposal-implementation-package/docs/proposal_for_regulation_of_ep_and_council_establishing_a_crisis_relocation_mechanism_en.pdf`.

European Commission. First report on relocation and resettlement, COM(2016) 165 final. *Communication from the Commission to the European Parliament, the European Council and the Council*, 2016a. URL `http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/european-agenda-migration/proposal-implementation-package/docs/20160316/first_report_on_relocation_and_resettlement_en.pdf`.

European Commission. Fifth report on relocation and resettlement. *Communication from the Commission to the European Parliament, the European Council and the Council*, 2016b. URL `http:`

//ec.europa.eu/dgs/home-affairs/what-we-do/policies/european-agenda-migration/proposal-implementation-package/docs/20160713/fifth_report_on_relocation_and_resettlement_en.pdf.

European Commission. Sixth report on relocation and resettlement. *Communication from the Commission to the European Parliament, the European Council and the Council*, 2016c. URL http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/european-agenda-migration/proposal-implementation-package/docs/20160928/sixth_report_on_relocation_and_resettlement_en.pdf.

European Commission. Member States Support to Emergency Relocation Mechanism (As of 21 July 2017). 2017. URL https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/european-agenda-migration/press-material/docs/state_of_play_-_relocation_en.pdf.

Eurostat. news release 217/2015, 2015. URL http://ec.europa.eu/eurostat/documents/2995521/7105334/3-10122015-AP-EN.pdf/04886524-58f2-40e9-995d-d97520e62a0e.

Eurostat. news release 44/2016, 2016. URL http://ec.europa.eu/eurostat/documents/2995521/7203832/3-04032016-AP-EN.pdf/790eba01-381c-4163-bcd2-a54959b99ed6.

Jesús Fernández-Huertas Moraga and Hillel Rapoport. Tradable Immigration Quotas. *Journal of Public Economics*, 115:94–108, 2014.

Jesús Fernández-Huertas Moraga and Hillel Rapoport. Tradable Refugee-admission Quotas and EU Asylum Policy. *CESifo Economic Studies*, 61 (3-4):638–672, 2015a.

Jesús Fernández-Huertas Moraga and Hillel Rapoport. Tradable Refugee-admission Quotas (TRAQs), the Syrian Crisis and the new European

Agenda on Migration. *IZA Journal of European Labor Studies*, (Discussion Paper No. 9418):1–16, 2015b.

Mark Fletcher. I'm a conservative, but this asylum seekers comic is disgusting. *AusOpinion, reprinted in The Guardian online*, 2014. URL `https://www.theguardian.com/commentisfree/2014/feb/13/asylum-seekers-graphic-campaign`.

Daniel Fragiadakis, Atsushi Iwasaki, Peter Troyan, Suguru Ueda, and Makoto Yokoo. Strategyproof Matching with Minimum Quotas. *ACM Transactions on Economics and Computation*, 4(1):6:1–6:40, 2015.

Richard B. Freeman. The Limits of Altruism: Selecting Living Donors. *American Medical Association Journal of Ethics*, 14(3):272–277, 2012.

Iason Gabriel. Effective Altruism and its Critics. *Journal of Applied Philosophy*, 34(4):457–473, 2017.

David Gale and Lloyd Shapley. College Admissions and the Stability of Marriage. *American Mathematical Monthly*, 69(1):9–15, 1962.

David Gauthier. *Morals by Agreement*. Clarendon, Oxford, 1987.

David Gauthier. How I learned to stop worrying and love the Prisoner's Dilemma. In Martin Peterson, editor, *The Prisoner's Dilemma*. Cambridge University Press, Cambridge, 2015.

Natalie Gold and Robert Sugden. Theories of Team Agency. In Fabienne Peter and Hans Bernhard Schmid, editors, *Rationality and Commitment*. Oxford University Press, Oxford, 2007.

Philip Grech. Undesired Effects of the European Commission's Refugee Distribution Key. *Preprint: ETH Zürich: Negotiation and Conflict Management Research Paper Series*, 2016.

Francesco Guala. Building Economic Machines: The FCC Auctions. *Studies in History and Philosophy of Science*, 32:453–77, 2001.

Francesco Guala. *The Methodology of Experimental Economics*. Cambridge University Press, Cambridge, 2005.

Francesco Guala. Getting the FCC auctions straight: a reply to Nik-Khah. *economic sociology newsletter*, 7(3):23–28, 2006.

Francesco Guala. How to Do Things with Experimental Economics. In Donald MacKenzie, Fabian Muniesa, and Lucia Siu, editors, *Do Economists Make Markets?*, pages 128–62. Princeton University Press, Princeton, 2007.

Francesco Guala. *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton University Press, Princeton and Oxford, 2016.

Francesco Guala. Précis of *Understanding Institutions*. *Philosophy of the Social Sciences*, 48(6):539–549, 2018a.

Francesco Guala. Solving the Hi-lo Paradox: Equilibria, Beliefs, and Coordination. In Anika Fiebich, editor, *Minimal Cooperation and Shared Agency*. Springer, 2018b.

Francesco Guala and Luigi Mittone. Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology*, 12(4):495–515, 2005.

John Halstead, Stefan Schubert, Mark Engelbert, and Hayden Wilkinson. Effective altruism: an elucidation and a defence. *Unpublished Manuscript*, 2019.

John C. Harsanyi. *Essays on Ethics, Social Behavior and Scientific Explanation*. D. Reidel Publ. Co., Dordrecht, 1976.

John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA, 1988.

Daniel Hausman. *The Inexact and Separate Science of Economics*. Cambridge University Press, Cambridge, 1992.

Friedrich A. Hayek. The Present State of the Debate. In F. A. Hayek, editor, *Collectivist Economic Planning*, pages 201–243. Routledge, London, 1935.

Friedrich A. Hayek. The Use of Knowledge in Society. *The American Economic Review*, 35(4):519–30, 1945.

Frank Hindriks and Francesco Guala. Institutions, rules, and equilibria: a unified theory. *Journal of Institutional Economics*, 11(3):459–480, 2015a.

Frank Hindriks and Francesco Guala. Understanding institutions: replies to Aoki, Binmore, Hodgson, Searle, Smith, and Sugden. *Journal of Institutional Economics*, 11(3):515–522, 2015b.

Bengt Holmström and Roger B. Myerson. Efficient and Durable Decision Rules with Incomplete Information. *Econometrica*, 51(6):1799–1819, 1983.

Joe Horton. The All or Nothing Problem. *The Journal of Philosophy*, 114(2): 94–104, 2017.

David Hume. Of the Independency of Parliament. In Eugene F. Miller, editor, *Essays, Moral, Political and Literary*. Liberty Fund, Inc. 1987, Cambridge, MA, 1742.

Susan Hurley. Social heuristics that make us smarter. *Philosophical Psychology*, 18:585–612, 2005a.

Susan Hurley. Rational Agency, Cooperation and Mind-reading. In N. Gold, editor, *Teamwork: Multi-Disciplinary Perspectives*, pages 200–215. Palgrave Macmillan, London, 2005b.

Leonid Hurwicz. Decentralized Resource Allocations. *Cowles Commission Discussion Paper*, Economics No. 2070:1–20, 1953.

Leonid Hurwicz. Optimality and Informational Efficiency in Resource Allocation Processes. In Kenneth J. Arrow, Samuel Karlin, and Patrick Suppes, editors, *Mathematical Methods in the Social Sciences*. Stanford University Press, Stanford, 1960.

Leonid Hurwicz. On informationally decentralized systems. In C. B. McGuire and R. Radner, editors, *Decision and Organization: a Volume in Honor of Jacob Marshak*, pages 297–336. North-Holland, Amsterdam, 1972.

Leonid Hurwicz. The Design of Mechanisms for Resource Allocation. *The American Economic Review*, 63(2):1–30, 1973.

Matthew O. Jackson. The Role of Theory in an Age of Design and Big Data. In Jean-Franois Laslier, Hervé Moulin, Remzi Sanver, and William S. Zwicker, editors, *The Future of Economic Design*. 2018.

Will Jones and Alexander Teytelboym. The International Refugee Match: A System that Respects Refugees' Preferences and the Priorities of States. *Refugee Survey Quarterly*, 36(2):84–109, 2017a.

Will Jones and Alexander Teytelboym. Matching Systems for Refugees. *Journal on Migration and Human Security*, 5(3):667–681, 2017b.

Will Jones and Alexander Teytelboym. The Local Refugee Match: Aligning Refugees' Preferences with the Capacities and Priorities of Localities. *Journal of Refugee Studies*, 31(2):152–178, 2018.

Jurgis Karpus and Natalie Gold. Team reasoning: theory and evidence. In J. Kiverstein, editor, *The Routledge Handbook of Philosophy of the Social Mind*, pages 400–417. Routledge Taylor Francis, Abingdon, 2017.

Jurgis Karpus and Mantas Radzvilas. Team Reasoning and a Measure of Mutual Advantage in Games. *Economics & Philosophy*, 34(1):1–30, 2018.

Bruce Katz, Luise Noring, and Nantke Garrelts. Cities and refugees: The German experience. *Brookings*, September 2016. URL `https://www.brookings.edu/research/cities-and-refugees-the-german-experience/?utm_campaign=Brookings+Brief&utm_source=hs_email&utm_medium=email&utm_content=34709340`.

Onur Kesten. School Choice with Consent. *Quarterly Journal of Economics*, 125:1297–1348, 2010.

Bettina Klaus and Flip Klijn. Procedurally fair and stable matching. *Economic Theory*, 27:431447, 2009.

Bettina Klaus, David F. Manlove, and Francesca Rossi. Matching under Preferences. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors, *Handbook of Computational Social Choice*, chapter 14, pages 333–355. Cambridge University Press, Cambridge, 2016.

Fuhito Kojima, Parag A. Pathak, and Alvin E. Roth. Matching with Couples: Stability and Incentives in Large Markets. *The Quarterly Journal of Economics*, 128(4):1585–1632, 2013.

Jaakko Kuorikoski and Aki Lehtinen. Incredible Worlds, Credible Results. *Erkenntnis*, 70(1):119–131, 2009.

Oskar Lange. Economics of socialism. *Journal of Political Economy*, 50(2): 299–303, 1942.

Abba Lerner. *The Economics of Control: Principles of Welfare Economics.* Macmillan, New York, 1944.

Shengwu Li. Ethics and Market Design. *Oxford Review of Economic Policy*, 33(4):70520, 2017a.

Shengwu Li. Obviously strategy-proof mechanisms. *American Economic Review*, 107(11):3257–3287, 2017b.

Leonie Lopp. *Regulations Regarding Living Organ Donation in Europe – Possibilities of Harmonisation.* Springer, Berlin, Heidelberg, 2013.

Matthias Lücke. The EU needs the mandate to fund and administer the asylum system. Technical report, GES (Global Economic Symposium), 2016.

William MacAskill. *Doing Good Better: How Effective Altruism Can Help You Make a Difference*. Random House, London, 2016.

David J C MacKay, Peter Cramton, Axel Ockenfels, and Steven Stoft. Price Carbon - I Will If You Will. In Peter Cramton, David J C MacKay, Axel Ockenfels, and Steven Stoft, editors, *Global Carbon Pricing: The Path to Climate Cooperation*. MIT Press, Cambridge, Massachusetts and London, England, 2017.

Uskali Mäki. MISSing the World. Models as Isolations and Credible Surrogate Systems. *Erkenntnis*, 70(1):29–43, 2009.

Uskali Mäki. Models and the locus of their truth. *Synthese*, 180(1):47–63, 2011.

Uskali Mäki. Modelling failure. In Hannes Leitgeb, Ilkka Niiniluoto, Päivi Seppälä, and Elliott Sober, editors, *Logic, Methodology and Philosophy of Science – Proceedings of the 15th International Congress (Helsinki)*, UK, 2017. College Publications.

Louis Makowski and Joseph M. Ostroy. General Equilibrium and Market Socialism: Clarifying the Logic of Competitive Markets. In P. Bardhan and J. Roemer, editors, *Market Socialism: The Current Debate*. Oxford University Press, Oxford, 1992.

Eric Maskin. Nash Equilibrium and Welfare Optimality. *The Review of Economic Studies*, 66(1):23–38, 1999.

Eric Maskin. Friedrich von Hayek and mechanism design. *Review of Austrian Economics*, 28:247–252, 2015.

R. Preston McAfee and John McMillan. Analyzing the Airwaves Auction. *Journal of Economic Perspectives*, 10(1):159–75, 1996.

Daniel McFadden. The human side of mechanism design: a tribute to Leo Hurwicz and Jean-Jacques Laffont. *Rev. Econ. Design*, 13:77–100, 2009.

Paul Milgrom. How obscure science led to spectrum auctions that connected the world. `http://thehill.com/blogs/pundits-blog/technology/331091-how-obscure-science-led-to-spectrum-auctions-that-connected-the`, 2017. Accessed: 2018-05-01.

David Miller. Immigrants, nations, and citizenship. *The Journal of Political Philosophy*, 16(4):371–390, 2008.

Robert A. Montgomery, Sommer E. Gentry, William H. Marks, Daniel S. Warren, Janet Hiller, Julie Houp, Andrea A. Zachary, J. Keith Melancon, Warren R. Maley, Hamid Rabb, Christopher Simpkins, and Dorry L. Segev. Domino paired kidney donation: a strategy to make best use of live non-directed donation. *Lancet*, 368:419–421, 2006.

Mary S. Morgan. Models, stories and the economic world. *Journal of Economic Methodology*, 8(3):361–384, 2001.

Mary S. Morgan. *The World in the Model: How Economists Work and Think*. Cambridge University Press, Cambridge, 2012.

Roger B. Myerson. Incentive Compatibility and the Bargaining Problem. *Econometrica*, 47:61–74, 1979.

Roger B. Myerson. Perspectives on Mechanism Design in Economic Theory. *The American Economic Review*, 98(3):586–603, 2008.

Roger B. Myerson. Fundamental theory of institutions: a lecture in honor of Leo Hurwicz. *Review of Economic Design*, 13(1):59–75, 2009.

Robert Northcott. The Efficiency Question in Economics. *Philosophy of Science*, 85:1140–1151, 2018.

International Communications Office. European Court Decision Won't Change Hungary's Immigration Policies. *Cabinet Office of the Prime Minister PM Orbán*, 2017. URL `http://abouthungary.hu/news-in-`

brief/pm-orban-european-court-decision-wont-change-hungarys-immigration-policies/.

Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. The MIT Press, Cambridge, Massachusetts, 1994.

Omar Parbhoo, Katy Davis, Robert Reynolds, Piyush Tantia, Pranav Trewn, and Sarah Welch. Best of Intentions: Using Behavioral Design to Unlock Charitable Giving. *ideas42 report*, 2018.

Derek Parfit. Equality and Priority. *Ratio*, 10:202–221, 1997.

Derek Parfit. Another Defence of the Priority View. *Utilitas*, 24:399–440, 2012.

Charles R. Plott. Experimental Methods in Political Economy: A Tool for Regulatory Research. In A.R. Ferguson, editor, *Attacking Regulatory Problems*, pages 117–43. Ballinger, Cambridge, Mass., 1981.

Eric A. Posner and E. Glen Weyl. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press, Princeton and Oxford, 2018.

Theron Pummer. Whether and Where to Give. *Philosophy & Public Affairs*, 44(1):77–95, 2016.

Wlodek Rabinowicz. Are Institutions Rules in Equilibrium? Comments on Guala's *Understanding Institutions*. *Philosophy of the Social Sciences*, 48 (6):569–584, 2018.

Felix T. Rapaport. The case for a living emotionally related international kidney donor exchange registry. *Transplantation Proceedings*, 18(3):5–9, 1986.

Hillel Rapoport. A fair and efficient European response to the refugee crisis, 2016. URL http://www.parisschoolofeconomics.eu/en/

economics-for-everyone/for-a-wider-audience/a-word-from/hillel-rapoport-a-fair-and-efficient-european-response-to-the-refugee-crisis/.

John Rawls. *A Theory of Justice*. Harvard University Press, Harvard, MA, 1971.

Michael A. Rees, Jonathan E. Kopke, Ronald P. Pelletier, Dorry L. Segev, Matthew E. Rutter, Alfredo J. Fabrega, Jeffrey Rogers, Oleh G. Pankewycz, Janet Hiller, Alvin E. Roth, Tuomas Sandholm, M. Utku Ünver, and Robert A. Montgomery. A Nonsimultaneous, Extended, Altruistic-Donor Chain. *The New England Journal of Medicine*, 360(11):1096–1101, 2009.

Michael A. Rees, T. B. Dunn, C. S. Kuhr, C. L. Marsh, Jeffrey Rogers, S. E. Rees, A. Cicero, L. J. Reece, Alvin E. Roth, O. Ekwenna, D. E. Fumo, K. D. Krawiec, Jonathan E. Kopke, S. Jain, M. Tan, and S. R. Paloyo. Kidney Exchange to Overcome Financial Barriers to Kidney Transplantation. *American Journal of Transplantation*, 17(3):782790, 2017.

Julian Reiss. The explanation paradox. *Journal of Economic Methodology*, 19 (1):43–62, 2012.

Andrew Rettman. Slovakia Filing Case against EU Migrant Relocation. *EU Observer*, (1), 2015. URL https://euobserver.com/justice/130499.

Mathew Robb, Chloe Brown, and Lisa Mumford. *Annual Report on Living Donor Kidney Transplantation. Report for 2017/18 (1 April 2003 – 31 March 2018)*. Statistics and Clinical Studies, NHS Blood and Transplant, 2018.

Alvin E. Roth. The Economics of Matching: Stability and Incentives. *Mathematics of Operations Research*, 7(4):617–28, 1982.

Alvin E. Roth. The Evolution of the Labor Market for Medical Interns and

Residents: A Case Study in Game Theory. *Journal of Political Economy*, 92:991–1016, 1984.

Alvin E. Roth. The College Admissions Problem is not Equivalent to the Marriage Problem. *Journal of Economic Theory*, 36:277–88, 1985.

Alvin E. Roth. The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics. *Econometrica*, 70(4): 1341–1378, 2002.

Alvin E. Roth. The Origins, History, and Design of the Resident Match. *Journal of the American Medical Association*, 289:909–12, 2003.

Alvin E. Roth. Repugnance as a Constraint on Markets. *Journal of Economic Perspectives*, 21(3):37–58, 2007.

Alvin E. Roth. What Have We Learned From Market Design? In Nir Vulkan, Alvin E. Roth, and Zvika Neeman, editors, *The Handbook of Market Design*. Oxford University Press, Oxford, 2013.

Alvin E. Roth. Transplantation: One Economist's Perspective. *Transplantation*, 99(2):261–264, 2015a.

Alvin E. Roth. *Who Gets What – And Why*. Houghton Mifflin Harcourt, Boston, 2015b.

Alvin E. Roth. Migrants aren't widgets. *Politico*, September 2015c. URL `http://www.politico.eu/article/migrants-arent-widgets-europe-eu-migrant-refugee-crisis/`.

Alvin E. Roth. Marketplaces, Markets, and Market Design. *American Economic Review*, 108(7):1609–1658, 2018.

Alvin E. Roth and Elliott Peranson. The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design. *American Economic Review*, 89(4):748–780, 1999.

Alvin E. Roth and Marilda A. Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge University Press, Cambridge, 1990.

Alvin E. Roth, Tayfun Sönmez, and M. Utku Ünver. Kidney Exchange. *The Quarterly Journal of Economics*, 119:457–488, 2004.

Alvin E. Roth, Tayfun Sönmez, and M. Utku Ünver. Pairwise kidney exchange. *Journal of Economic Theory*, 125:151–188, 2005.

Alvin E. Roth, Tayfun Sönmez, M. Utku Ünver, Frank L. Delmonico, and Susan L. Saidman. Utilizing list exchange and nondirected donation through 'chain' paired kidney donations. *American Journal of Transplantation*, 6: 2694–2705, 2006.

Alvin E. Roth, Tayfun Sönmez, and M. Utku Ünver. Efficient Kidney Exchange: Coincidence of Wants in Markets with Compatibility-Based Preferences. *The American Economic Review*, 97(3):828–851, 2007.

Alvin E. Roth, Tayfun Sönmez, and M. Utku Ünver. Kidney paired donation with compatible pairs. *American Journal of Transplantation*, 8:463, 2008.

Ariel Rubinstein. Dilemmas of an Economic Theorist. *Econometrica*, 74(4): 865–883, 2006.

Ariel Rubinstein. *Economic Fables*. Open Book Publishers, Cambridge, UK, 2012.

Michael J. Sandel. *What Money Can't Buy: The Moral Limits of Markets*. Farrar, Straus and Giroux, New York, 2012.

Thomas Schelling. Game Theory: a Practitioner's Approach. *Economics and Philosophy*, 26:27–46, 2010.

DL Segev, AD Muzaale, BS Caffo, and et al. Perioperative mortality and long-term survival following live kidney donation. *Journal of the American Medical Association*, 303(10):959–966, 2010.

Amartya K. Sen. Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs*, 6(4):317–344, 1977.

Andrew E. Shacknove. Who Is a Refugee? *Ethics*, 95(2):274–284, 1985.

Peter Singer. Famine, Affluence, and Morality. *Philosophy & Public Affairs*, 1(3):229–243, 1972.

Peter Singer. *The Most Good You Can Do: How Effective Altruism Is Changing Ideas About Living Ethically.* Yale University Press, New Haven and London, 2015.

Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations.* W. Straham and T. Cadell, London, 1776.

Tayfun Sönmez. Manipulation via Capacities in Two-Sided Matching Markets. *Journal of Economic Theory*, 77(1):197–204, November 1997.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* MIT Press, 2000.

Bureau of Labor Statistics. National Census of Fatal Occupational Injuries in 2016. *News Release*, 2017.

Robert Sugden. Thinking as a Team: Towards an Explanation of Nonselfish Behavior. *Social Philosophy and Policy*, 10(1):69–89, 1993.

Robert Sugden. Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology*, 7:1–31, 2000.

Robert Sugden. The Logic of Team Reasoning. *Philosophical Explorations*, 6 (3):165–181, 2003.

Robert Sugden. Mutual advantage, conventions and team reasoning. *International Review of Economics*, 58:9–20, 2011.

Robert Sugden. Team Reasoning and Intentional Cooperation for Mutual Benefit. *Journal of Social Ontology*, 1(1):143–166, 2015.

Evelyn M. Tenenbaum. Bartering for a Compatible Kidney Using Your Incompatible, Live Kidney Donor: Legal and Ethical Issues Related to Kidney Chains. *American Journal of Law & Medicine*, 42:129–169, 2016.

The Royal Swedish Academy of Sciences. The Prize in Economic Sciences 2007, 2007. URL `https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2007/popular-economicsciences2007.pdf`.

The Royal Swedish Academy of Sciences. The Prize in Economic Sciences 2012, 2012. URL `https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2012/press_02.pdf`.

Walter Trockel. Integrating the Nash program into mechanism theory. *Review of Economic Design*, 7(1):27–43, 2002.

UNHCR. Hungary violating international law in response to migration crisis: Zeid, 2015. URL `http://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=16449&LangID=E`.

UNHCR. Projected Global Resettlement Needs 2017, 2016. URL `http://reliefweb.int/sites/reliefweb.int/files/resources/575836267.pdf`.

Philippe van Basshuysen. The Prisoner's Dilemma. *Economics and Philosophy*, 33(1):153–160, 2016.

Philippe van Basshuysen. Towards a Fair Distribution Mechanism for Asylum. *Games*, 8(4):41, 2017.

Philippe van Basshuysen. Radical Markets: Uprooting Capitalism and Democracy for a Just Society. *Review of Political Economy*, forthcoming.

Michael G. Vann. Of Rats, Rice, and Race: The Great Hanoi Rat Massacre, an Episode in French Colonial History. *French Colonial History*, 4:191–203, 2003.

William S. Vickrey. Counterspeculation, Auctions, and Competitive Sealed Tenders. *The Journal of Finance*, 16(1):8–37, 1961.

Ludwig von Mises. Economic Calculation in the Socialist Commonwealth. In F. A. Hayek, editor, *Collectivist Economic Planning*, pages 87–130. Routledge, London, 1935.

C. Bradley Wallis, Kannan P. Samy, Alvin E. Roth, and Michael A. Rees. Kidney paired donation. *Nephrol Dial Transplant*, 26:2091–2099, 2011.

Haidong Wang and et al. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 19802015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388:1459–1544, 2016.

Wissenschaftlicher Dienst. Die Cross-over-Lebendspende. Zum Stand in Deutschland und in ausgewählten europäischen Ländern. *Deutscher Bundestag*, WD9-3000-022/17, 2017.

E. Steve Woodle, J.A. Daller, M. Aeder, R. Shapiro, T. Sandholm, V. Casingal, D. Goldfarb, R.M. Lewis, J. Goebel, and M. Siegler. Ethical Considerations for Participation of Nondirected Living Donors in Kidney Exchange Programs. *American Journal of Transplantation*, 10:1460–1467, 2010.

James Woodward. *Making Things Happen*. Oxford University Press, Oxford, 2003.