How does game theory inform economic engineering?

Philippe van Basshuysen

April 13, 2018

Abstract

How is it possible that models from game theory, which are typically highly idealised, can be harnessed for designing institutions through which we interact? I argue that game theory assumes that social interactions have a specific structure, which is uncovered with the help of directed graphs. The graphs make explicit how game theory encodes counterfactual information in natural collections of its models and can therefore be used to track how model-interventions change model-outcomes. For model-interventions to inform real-world design requires the truth of a causal hypothesis, namely that structural relations specified in a model approximate causal relations in the target interaction; or in other words, that the directed graph can be interpreted causally. In order to increase their confidence in this hypothesis, market designers complement their models with natural and laboratory experiments, and computational methods. Throughout the paper, the reform of a matching market for medical residents provides a case study for my proposed view, which hasn't been previously considered in the philosophy of science.

Keywords Economic engineering, market design, matching theory, NRMP, game theory, models, causal graphs.

1 Introduction

While economists often take market outcomes as phenomena to be explained or predicted, recent years have seen an increasing interest in *market design*. Instead of taking them as given, economists in this field seek to bring about desirable outcomes by actively shaping, or "engineering" markets. This practice poses interesting new challenges for the philosophy of science.[1] This paper is concerned with such a challenge: the way that economists qua market

[1] Market design also raises questions for moral philosophy, viz., according to what criteria market outcomes should be considered desirable. See Li [2017] for a general treatment of ethics and market design, and van Basshuysen [2017] for the ethics of a particular market design, namely a matching market for asylum. Here, I take criteria to be exogenously given through policy goals.

designers harness economic models.

While there has been a large debate among philosophers of science and economic theorists on how economic models relate to their intended targets,[2] this question takes an interesting twist in the context of market design. One reason is that, since designers do not only model markets but use these models to change the "rules" that govern those markets, this practice directs attention to the way in which counterfactual information is encoded in the models used. What can a model tell us about how a market outcome would change if the rules of the market changed, and how can this information be used for market design? Second, in the debate, it has been emphasised that economic models depend crucially on false assumptions, and isolate causal mechanisms that are in the real world interfered with by distorting mechanisms, thus in the best case representing stylised facts. But market design requires confidence in the prediction that a design will perform well in the messy real world. How can models, possibly in combination with other tools, establish such confidence?

I shall attempt to answer these questions with regards to an important class of economic models, namely models from game theory. It is frequently emphasised that game theory plays a vital role for design purposes. For example, Roth and Sotomayor [1990] write,

> "It is this close observed connection between individual incentives and market behavior that suggests that, however game theory may need to be further developed as a descriptive theory, it has a critical role to play in helping us to understand and design the institutions through which we interact."[3]

I shall take an institution as a case study, which hasn't been previously considered in the philosophy of science: the matching market that allows medical graduates in the US to find training positions in public hospitals. In this market, the graduates state their preferences over hospitals and the hospitals state their preferences over graduates, and an algorithm is used to determine the matchings relative to the preferences submitted. In the 1990s, game theorists were significantly involved in the design of a new algorithm which was commissioned as a response to market failure.

How is it possible that game theoretic (GT) models informed the successful reform of this market? I shall argue that these models assume that social interactions have a specific structure.

[2] Some examples are the following: Alexandrova [2006a, 2008], Alexandrova and Northcott [2009, 2015], Cartwright [1989, 1999], Hausman [1992, 2015], Mäki [2011, 2017], Morgan [2001, 2002], Reiss [2012], Rodrik [2015], Roth [1991, 2002], Rubinstein [2006, 2012], Schelling [2010], Spiegler [2015], Sugden [2000], Sutton [2002], Ylikoski and Aydinonat [2014]. Moreover, *Erkenntnis* devoted a special issue, "Economic Models as Credible Worlds or as Isolating Tools?" (2009), to this debate.

[3] It seems a fair interpretation that with "institution", Roth and Sotomayor just mean "market" in this quote, since it stems from a classic text on matching theory, i.e. the study of *matching markets* (see section 5). An alternative reading is that institutions include markets in the sense that markets are institutions but not all institutions are markets. Either interpretation fits my purpose to emphasise the importance of game theory for the design of markets.

One of the contributions of this paper is to provide an account of GT models, which uncovers this structure. I will show that directed graphs can be lifted quite naturally out of standard game-theoretic practice, and the graphs reveal that game theory encodes counterfactual information in natural collections of its models.

Once this has been done, graphs provide a particularly useful analysis from the perspective of policy-making. If the graph of the model can be interpreted as a *causal graph*, which approximately describes a causal mechanism in effect in the target interaction, then the model-interventions do accurately describe how interventions in the real world would change outcomes, and this fact can be used to design markets.

To interpret the directed graph as a causal graph presupposes the truth of a causal hypothesis: that the model represents a relevant causal mechanism, which is not significantly distorted by other mechanisms. But what provides confidence in the truth of this hypothesis, given what has been said above, that economic models make false assumptions and isolate mechanisms which are typically interfered with by other mechanisms? Another look at the case study is revealing: it shows that, in order to confirm causal hypotheses, the models used are in a specific way complemented with natural, laboratory, and computational experiments.

This paper is organised as follows. In section (2), I introduce my case study, the matching market for medical residents. In section (3), I put forward a directed graph of GT models. The graph makes explicit how game theory encodes counterfactual information in collections of models, a fact that explains the methodology of market design, which is object of section (4). But the methodology poses the puzzle of how inferences from GT models to the real world are established. In section (5), the case study is revisited to show how practitioners grapple with this question, and what factors increase their confidence that their models would indeed successfully guide the reform of the market. Section (6) discusses the lessons learned from the case study within my account of GT models, and connects them to some of the literature on economic models. Section (7) concludes.

2 The matching market for medical residents

The National Resident Matching Program (NRMP) is a clearinghouse that matches medical graduates ("residents") with training positions ("residencies") in public hospitals in the US. Residencies allow graduates to specialise in a specific medical branch, and they are a requirement for practising as a physician. Residents and hospitals both submit preferences to the NRMP, which uses an algorithm to determine the matchings. In the 1990s, the algorithm in use had generated considerable discontent of prospective residents. The game theorist and later Nobel laureate Alvin Roth was commissioned to direct the design of a new matching algorithm. Since this important case of market design has to date not been considered by philosophers of science, I will describe the history and the reform of the market in some detail, based on Roth's and his collaborators' accounts.[4]

Before the NRMP was established, the labour market for medical interns was a decentralised market that had been subject to various market failures. For example, hospitals offered positions to students in ever earlier stages of their medical studies in order to prevent other hospitals from "snatching" interns first. By 1940, students were offered positions up to two years before graduation, which was prejudicial for both hospitals and students: hospitals were facing high uncertainty about the future performance in medical school of the students they hired, and students were left under pressure to make career choices very early.

This situation changed after a reform had been implemented, which prescribed a maximum time period before students' graduation that the hospitals had to adhere to before offering internships. This, however, led to different inefficiencies: when offers for internships were issued, hospitals demanded that offer holders would accept offers in ever shorter time periods in order to render it impossible for them to wait for preferred offers; and hospitals that didn't send out sufficiently many offers sufficiently early would frequently send out offers to students already in a different internship – even though they may have preferred the later offer.

Eventually, in 1951, the NRMP was introduced as a response to the market failures. It collects rank order lists ("ROLs"): lists that reflect the students' preferences over the hospitals they had previously had an interview with, and the hospitals' preferences over the students they had interviewed. The assignments are then determined algorithmically. Since students and hospitals are free to decide whether to find residencies and residents, respectively, through the centralised clearinghouse or on their own, high rates of voluntary participation in the system in its early years (over 95 %) can be interpreted as evidence for the well-functioning of the market and the satisfaction of the agents participating in it.

However, over the years changes occurred in the market, which would eventually result in a crisis of confidence of the applicants in the market. An example of such a change is that 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.[5] Members of couples often have interrelated preferences, particularly to find positions

[4] Roth's interest in the market dates back at least to Roth [1982]. Technical or historical accounts can be found in Roth [1984], Roth and Sotomayor [1990], Roth and Peranson [1999], Roth [2002, 2003, 2013, 2015], Kojima et al. [2013], amongst others.

[5] The integration of couples is only one of four types of "match variations" (Roth and Peranson [1999]) and the only that I shall discuss. Other examples are hospitals with interlinked numbers of positions such as, say, five in the neurology department if internal medicine fills all its positions, fewer otherwise.

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.

The initial algorithm could not accommodate such desires because it would process only single preference lists. The NRMP modified the system to permit couples to hand in a pair of ROLs together and to specify a “leading member”. The algorithm would then match the leading member first and next the preference list of the other member would be edited to eliminate positions far from that of the leading member. However, this rather ad-hoc modification of the system could not prevent rates of participation from dropping.

The accommodation of couples in the system was not the only challenge the NRMP was facing. Another one was that the numbers of students relative to residencies offered increased substantially over the years,[6] which led to matchings being less favourable for students. And more examples could be added. In the 1990s, the dissatisfaction among applicants – as expressed by various student associations – was at a peak. Many claimed that the algorithm would show favouritism to the hospitals at the expense of the graduates; 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 algorithm that was used to determine the matchings, 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 algorithm for conducting the matchings. The goal of the design was an algorithm that would remove agents’ incentives to make arrangements outside the system. Moreover, the matchings should be as favourable as possible for applicants, while keeping to a minimum the possibility of strategic behaviour. Roth directed the design of the algorithm which is now known as the “Roth-Peranson algorithm”.[7] It was first implemented in 1998, has been working successfully since, and has been adopted by numerous labour market clearinghouses. The question I seek to answer is, how did GT models inform the successful reform of the matching market?

3 Making game theoretic counterfactuals explicit

Before I present my account of GT models, two general clarifications are in order: one concerning what GT models can be used for, and one on the level of description at which I take them to

[6] In the year that the algorithm was introduced, the number of internships offered was almost twice the number of residents who could take them. This changed partly because foreign medical residents were allowed to participate in the matching programme in the 70s (Roth [1984]).

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

operate. In the most general terms, in social interactions such as markets, individuals face choices which collectively result in outcomes with certain features (Buchanan [2001]). In our case study, individuals' choices are to hand in ROLs (or not to take part in the centralised market), and the outcome is a matching with certain features, e.g. Pareto-efficiency with respect to the ROLs submitted. This leads me to my first clarification: rather than understand a particular matching in a particular year, we are interested in a general feature of a matching market, e.g. what market failures are likely to occur in markets that share a certain structure. An institutional design must be reliable in the sense that, even if it is to organise a one-shot interaction (as would be the case if the NRMP were to clear only once), at least in principle it should be possible to repeat the interaction and achieve a similar outcome. I take it that the level of description that GT models adopt is what types of social interactions result in what types of outcomes. Accordingly, in the following I shall mean types of events with "outcomes", "choices", etc.

Second, to explain outcomes, which eventually makes policy-making possible, requires understanding two things: how outcomes follow from combinations of agents' choices, and what brings about those choices in the first place. The former constitutes the rules of the market which are usually publicly observable. The latter includes agents' private information: their beliefs, desires and their reasoning processes, which together constitute their *motivations*, or *incentives* to make choices.[8] The two are intertwined: the rules of a situation influence agents' incentives to act in certain ways – which in turn may make it desirable for a planner to change the rules of the situation. Game theory operates where the two domains, the private and the public, interact.[9] This will be a recurrent theme in the following, and it should give a prima facie reason for why models from game theory are relevant to the case study.

But what are GT models? In my proposal, they are formal structures of sets and relations between them. To acquire meaning, interpretations connect the formal structures to target interactions. 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 they result in choices, as well as which outcomes result from combinations of choices. Thus, they include things in both the private and the public domains.

Market designers treat institutions as variables to be intervened on (cf. Guala [2007]). I take this view literally, and I shall argue that the structure which GT models impose on the defined

[8] Both terms can be found in the literature and I use them synonymously.

[9] Cf. "Game theory seeks to understand economic environments by analysing how the motivations of the agents interact with the 'rules of the game' – that is, the customs, rules, procedures, and constraints around which a market may be organized" (Roth and Sotomayor [1990][p. 10]).

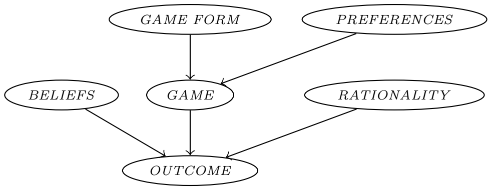

Figure 1: Graph of GT models. The nodes in the graph are variables and edges represent functional relations.

entities can be described by directed graphs, as in Figure 1. The graph consists in a set of nodes and directed edges between the nodes. As is standard in the literature on directed graphs, call node X the parent of node Y and Y the child of X if there is an edge from X to Y. Call nodes without children, leaf nodes, and nodes without parents, root 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 random variables whenever this does not cause confusion. The value of a given child is a function of the values of its parents. For example, the graph specifies that *OUTCOME* is a function of *GAME*, *BELIEFS*, and *RATIONALITY*.

Informally, the graph can be described as follows. Starting from the top, there is a *GAME FORM* variable that takes as value the "rules" that govern an interaction and a *PREFERENCES* variable that takes as value the players' preference relations. Together, these variables define the value of a *GAME* variable: a particular game, for example, a Prisoners' Dilemma in normal form. Together with a *RATIONALITY* and a *BELIEFS* variable, *GAME* defines the outcome, or set of outcomes, of the game. This is specified by the *OUTCOME* variable which is the only leaf node in the graph.

What justifies my claim that, if a GT model accurately represents a social interaction, then so does my graph of it? I take it to indirectly confirm my account that it will shed light on the practice of market design. For now, I shall show that the graph follows naturally from the way in which game theory is taught. Standard textbooks (e.g. Osborne and Rubinstein [1994]) usually 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 $\mathbf{A} = \times_{i \in N} A_i$.[10] For now, think of the *GAME* variable in the graph as ranging over different games Γ in the mathematical sense above. 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.

In textbooks, the definition of a game is usually followed by the introduction of solution con-

[10] For ease of exposition I shall ignore mixed strategies here and in the following. 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 $i \in N$.

cepts, 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. In the abstract, 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 (cf. Aumann and Brandenburger [1995]). 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*, *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 action profiles but over consequences of action profiles. This case can be accommodated if we add to the game a set C of consequences and $g : \mathbf{A} \longrightarrow C$ a function from action profiles to consequences. The preferences $(\succeq_i)_{i \in N}$ range over C. So in this general case which we shall operate with here, 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 define two separate variables: 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 "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).

The following example uses a simple game to illustrate the structure of GT models.

Example 1 (Prisoners' Dilemma and Prisoners' Delight)**.** *Suppose GAME FORM takes as value the interaction specified in Table 1. There are two players, Row and Col, who can both cooperate or defect. The four possible action profiles result in consequences a, b, c, and d.*

$GAMEFORM =$

	cooperate	defect
cooperate	a	b
defect	c	d

Table 1: *GAME FORM* takes as value 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 specified in the table. Together with the players' preferences, this game form determines a game.

$GAMEFORM' =$

	cooperate	defect
cooperate	a	c
defect	b	d

Table 2: A different value of *GAME FORM* as a result of a tax on defection and subsidy on cooperation.

Suppose 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 players 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 defect is a strictly dominant strategy for both players.) Then, both players will defect and OUTCOME takes value d.*

Now, suppose that the value of GAME FORM changes as specified in Table 2. The new game form could be the result of a change of an institutional rule, e.g. a tax on defection against cooperation, or a subsidy on the converse behaviour.

This 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 changes from d to 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 to calculate what the corresponding outcomes are for different values of variables. This is not so visible in the standard GT language, but the directed graph proposal provides a way to make this property explicit. The counterfactual information encoded in a model is at the core of my argument of how game theory can be harnessed for market design. The intuition is that designers use this information to intervene on the *GAME FORM* variable: game forms are designed which "force" players to produce outcomes that are considered desirable.

Before moving on to a more thorough consideration of market design, I should hasten to delineate the scope of the graph in Figure 1. Game theory is a family of diverse models, and it must be made clear which class of models the graph is intended to capture. Two points of clarification are required here. First, the graph is intended primarily to capture the structure of one-shot games in normal form. Extensive games can be used to additionally model social interactions where players' actions are sequential. In such interactions, players update their beliefs as the play unrolls. Extensive games can capture such additional information, which makes possible more refined solution concepts such as subgame perfect equilibrium. While I see no reason that extensive games cannot be captured by extending the graph to allow players to update beliefs, this is not within the scope of this paper.

Second, while extensive games contain information which the graph is not intended to capture, *cooperative games* get by with less information than games in normal form, and by suppressing some information, the graph can capture these games. This is important to note because market designers frequently exert tools from both cooperative and noncooperative game theory (the matching market is a case in point). In the cooperative case, the $GAME$ variable ranges over cooperative games. What distinguishes the two kinds of games is the different levels of abstraction they assume: whereas non-cooperative game theory aims at giving a detailed description of the situation, namely the players' actions that result in outcomes, cooperative game theory suppresses information regarding actions, and characterises the rules of the game through the payoffs that players and coalitions of players can achieve.[11] Moreover, cooperative game theory assumes that outcomes are enforceable. Instead of following from assumptions on agents' rationality and beliefs, cooperative solution concepts are characterised axiomatically. Thus, a cooperative solution concept determines the value of the $OUTCOME$ variable as a function of the $GAME$ variable alone. The values of the $RATIONALITY$ and $BELIEFS$ variables are not specified. I call a graph in which the values of some variables are not specified, an incomplete graph. The graph of a cooperative game is incomplete in this sense.

The relation between the noncooperative and the cooperative theory is provided by the so-called Nash programme (cf. Nash [1953]). It asks to specify noncooperative models whose equilibria coincide with cooperative solutions. For example, the core, one of the most widely used solution concepts in cooperative game theory, can be implemented through a variety of noncooperative procedures of proposal-making, given that some anonymity conditions are satisfied (cf. Serrano [2005]). In terms of graphs, what happens in a specific result of the Nash programme is that the incomplete graph of a cooperative game is extended to a complete, noncooperative one by adding players' actions in the $GAME$ variable, and specifying values for the $BELIEFS$ and $RATIONALITY$ variables. The value of the OUTCOME variable remains the

[11] A simple example of a cooperative game is a coalitional form game with transferable utility, which is defined as a pair $< N, v >$ where N is a set of players and v a characteristic function that specifies, for each non-empty subset S of N, the payoff profiles feasible for S. The theory then offers a variety of solution concepts, each a function that associates any game $< N, v >$ with a payoff vector that specifies each player's payoff. In this case, the GAME variable would range over $< N, v >$.

same but as a consequence of the agents' strategic actions.

4 From model-interventions to real world interventions: market design, and challenges from the philosophy of science

Market designers seek to realise outcomes that possess properties which a policy maker considers desirable. They proceed by imposing institutional rules which exploit agents' rational pursuit of their individual goals. To predict how agents' behaviour would adapt after a rule is changed, they model social interactions as games. Within the model, different games can be compared on the basis of how close their associated equilibria are to the policy goals in question, or more precisely, to what are interpreted to be the policy goals within the model.

This practice can be helpfully clarified using the directed graph model of the previous section. My proposal is that we view game forms as variables, whose values may determine particular games. The policy-maker may then view their choice as over possible values of the game form, each of which prescribes particular possible games, which they may ultimately impose on the social system. Since game forms induce games, the former can be compared in lieu of the games directly, with the goal of eventually imposing the game form chosen as a policy. In terms of our graphical representation, the *GAME FORM*[12] variable is the variable where *model-interventions* take place: different game forms are compared by calculating the equilibria of their associated games and the extent to which they conform to their policy goals.

However, this methodology faces a challenge: in order to know what game is induced by a game form, the value of the *PREFERENCES* variable must normally be specified. But the designer does not know the agents' preferences, and they may have incentives to obscure their preferences so 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. Furthermore, in a market such as the NRMP which clears regularly, it may not be known to the designer in advance how many agents there will be in the market. In short, the challenge is that a game form may induce different games for different populations and different types of players.

Market designers seek to overcome this problem by looking at variable types and populations

[12] In the context of design, game forms are called *mechanisms* – hence the term "mechanism design" for the part of market design which is concerned with inventing mechanisms that implement social choice functions. I do not adopt this terminology because I reserve the term "mechanism" for an explicitly causal meaning (see below). This choice of terms is closer to standard philosophy of science terminology.

of players. A game form is said to implement a property according to a solution concept if for any possible combination of players and types, there is an action profile that is prescribed by the solution concept, and the corresponding outcome satisfies the property. The game forms of interest to the designer implement desirable properties, a methodology that I will interpret as systematic robustness analysis below. The final step of the design is to select the game form that is closest to implementing properties that are interpreted as the policy goals within the model. If the design is successful, the rules which correspond to the chosen game form in the model induce the intended policy goals in the real world.

The "if successful"-part in the final step poses a puzzle: what provides confidence in the prediction that a design will perform as intended in the real world? A prediction of this kind seems to be based on the hypothesis, roughly, that interventions within the model inform interventions in the real world in the sense that the latter interventions reliably establish outcomes that resemble those in the model. So it would seem that what is presupposed is a *causal hypothesis*. It can be split into two parts: (i) the structure of the model corresponds to a stable causal mechanism in the real world; and (ii) the game form in the model implements properties that correspond to the relevant policy goals. The truth of (i) and (ii) would justify confidence in the prediction that a policy that corresponds 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 is a *causal graph* of its target interaction. In a causal graph, the variables range over events in the real world, and the edges indicate causal relationships.[13] If the graph is a causal graph of the interaction that is being modelled, the values of the variables are specific events which together constitute the interaction. For example, the variable *GAME FORM* takes as value the event that a specific rule governs the target interaction, that is, which choices are available to agents and how their choices jointly result in an outcome. 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.

In a causal graph, directed edges indicate *direct causation*:[14] 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

[13] To formally show that it is in principle possible to interpret the graph in Figure 1 as a causal graph in the sense of, e.g. Spirtes et al. [2000], we must specify a probability distribution over the graph which satisfies the causal Markov condition and the minimality condition. These axioms are trivially satisfied by the standard GT textbook interpretation of the nodes that I gave above.

[14] 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 my account particularly well.

the causal hypothesis is true and the graph can be interpreted causally, the structural relations in the model describe a causal mechanism effective in the real-world interaction. Moreover, the hypothetical interventions in the model reveal information about how the outcome of the target interaction would change as the rules that govern it change.

But how is the causal hypothesis justified, the truth of which would license the epistemic step from the model to the real world? What establishes confidence that we can treat the graph as a causal graph? An answer to this question would help explain how some serious and well-known challenges from the philosophy of science are met in the context of market design. These challenges have been presented eloquently by Nancy Cartwright, among others. In Cartwright [2009], she argues that typical models in economics face two problems. First, they are overconstrained, i.e. they depend crucially on false assumptions. These assumptions cannot be relaxed without considerably weakening the result of the model. In terms of our graphical representation, the values of the variables in an overconstrained model do not faithfully represent events that constitute the target situation. For example, the *BELIEFS* variable may take the value that players have correct beliefs about their opponents' preferences in the model, but they do not in the target interaction. But relaxing the assumption of correct beliefs, the model may be too weak to identify equilibria. So we lack confidence that agents act according to our predictions based on equilibria.

Second, in the causal hypothesis, it is implicitly assumed that the causal mechanism which the model describes is the only mechanism in place. In Cartwright [1989]'s words, models "isolate capacities". (Uskali Mäki holds similar views on models as isolating tools, e.g. Mäki [2009, 2011].) But social phenomena are complex, and there may be interfering mechanisms in the real world that distort the capacity which the model isolates (Cartwright [1999]). In terms of the graph, interfering mechanisms would mean that it lacks nodes or edges that are relevant in the target interaction, that is, that may change the outcome. Cartwright [2009] argues that, even if these other mechanisms could be isolated too, economics typically lacks laws of composition that would allow to faithfully describe how these causal mechanisms interact in the target interaction. Furthermore, it may be in principle impossible to complete a graph. For instance, there may be capacities that remain invisible because they are blocked by structural circumstances from exhibiting themselves. Think of a person's capacity for intellectual curiosity, independence of thought, etc., in a state of malnutrition.

Before I attempt to give an answer to these challenges, it is now the time to reconsider our case study, which will show how market designers grapple with these questions, and the way they do so will hint at an answer to the questions.

5 The matching market for medical residents revisited

Let's briefly recall the case of the NRMP. In the 1990s, the market was subject to various market failures: rates of participation were low, and applicants complained that the algorithm in use to determine the matchings would show favouritism to the hospitals at their expense, and that it would pay to "trick the system" by handing in ROLs strategically. The NRMP directors commissioned a new matching algorithm to alleviate the market from these failures. They set three policy goals which the new algorithm should implement as far as possible: to give the agents incentives to stick to the matchings, to make the resulting matchings as favourable as possible for the applicants, and to reduce their opportunities for strategic behaviour. In the following, I shall sketch a simple GT model of the market – more precisely, a model from a subdiscipline of game theory called *matching theory* – and some of the theoretical results that hold in this model. I shall then flesh out three lessons about how the model was manipulated, enriched, and complemented with other tools to inform the reform of the market.

From a game theory perspective, a matching algorithm functions as a game form that defines, for the agents' – that is, the applicants' and hospitals' – preferences, 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 specific 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 that she had an interview with and that she finds acceptable, and for each hospital over the students that it had interviewed and that it finds acceptable.[15] The agents' actions are to submit ROLs. Formally, 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 algorithm 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 algorithm in use by the time the NRMP directors commissioned the new design. As shown in Roth [1984], this

[15] The hospitals' preference lists underdetermine possible preferences the hospitals may have over *groups of residents* (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). It is usually assumed that hospitals' preferences over residents are *responsive* (cf. Roth [1985]): roughly, hospitals 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 and preferred to s_j. This is a false assumption of the kind that will be discussed in the next section.

algorithm is in the simple model described above equivalent to the *hospital-proposing deferred acceptance algorithm* (DAA^H). It therefore suffices to sketch the latter algorithm:

- In the first step, each hospital "proposes"[16] to the highest-ranked students 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 and 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 paper Gale and Shapley [1962], in the simple model described above, this algorithm 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. 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 higher ranked on a hospital h_i's ROL than a student matched to it, it must have applied to s_j at a previous step and been rejected. So s_j must have ranked h_i lower than her actual match and (s_j, h_i) is not a blocking pair.

Note that 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 GT models, and consequently, it was not known to be stable in this sense. 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 suggests naturally that stability is the formal equivalent to the directors' first policy goal to provide incentives to stick to the matchings.

However, this is a hypothesis on the basis of the model alone. The designers required stronger evidence. To gain confidence that stability really does achieve this goal, they resorted to natural experiments. Roth [2002] compares regional matching markets for physicians and

[16] Since Gale and Shapley [1962]'s first statement of the algorithm in the context of the "Marriage Market", it has been common to describe the algorithm using the predicates 'propose' and 'accept'/'reject'. This is for ease of presentation; what is of course meant is not that the agents act in a decentralised market but the central processing of the ROLs by the NRMP (Roth [2002]).

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. The two remaining markets used stable algorithms, and both were performing well. This gave evidence to the importance of stability. Now, it is logically possible that the survival or not of the different markets is 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 increasing the evidence that stability is key for achieving the first goal stated by the directors.[17] I take this to be the first lesson from the case study: *a simple model suggested a property which seemed to correspond to a policy goal, and a game form that implements that property. Natural and laboratory experiments were used to provide evidence for this hypothesis.*

What about the other policy goals of the directors? The simple model above can be used to substitute different algorithms for DAA^H, and compare the outcomes with respect to the policy goals. Take DAA^S, which is the algorithm equivalent to DAA^H but with the roles of the students and the hospitals switched. As shown in Gale and Shapley [1962], DAA^S produces matchings that are stable too, but which for all students weakly Pareto-dominate any other stable matchings with respect to their ROLs submitted, whereas for the hospitals, the matchings from DAA^H weakly Pareto-dominate all stable matchings. Since a goal of the directors was to produce stable matchings as favourable as possible for students, we may hypothesise that DAA^S performs better on this front. Moreover, DAA^S makes it a dominant strategy for all the students to submit their true preferences. This does not hold for all hospitals. But it seems that strategic behaviour in the market was particularly conspicuous on the part of the students, and banning this possibility was seen as a priority. Moreover, 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). Furthermore, there will be some room for strategic behaviour because there is no stable algorithm that makes it a dominant strategy for all agents to reveal their preferences. So DAA^S seems to be the algorithm of choice, if stability is taken as a first-order goal.

However, the above model turns out to be somewhat too naive for design purposes because the market possesses relevant features that the model does not reflect. 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 described above. But once the model is extended to incorporate couples, some of the theorems described above do not generalise to this case;

[17] A follow up experiment, mentioned in Roth [2002], added an explanation of why the two unstable markets survived: this was less due to the unstable algorithms in use but rather due to features of these markets that the other markets lacked – e.g., they were much smaller.

in particular, as shown in Roth [1984], the set of stable matchings may be empty. This is the second lesson from the NRMP: *negative results were important. They revealed which features of the market that the model fails to represent may be relevant.*

So the designers knew from simple models in combination with experiments that stability is key, but that stability cannot be guaranteed in the target market. This is not the end of the matter. They next asked whether there is 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 decreased.[18] Roth and Peranson [1999] designed a modified DAA^S which detects blocking pairs and repairs them, if possible, at intermediate steps. So this algorithm seeks to find stable matchings.

The Roth-Peranson algorithm is much more complex than a simple DAA^S (for an insightful graphical representation of the algorithm, cf. 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.

These were not the only computational experiments. The impossibility theorem above said that the set of stable matchings could be empty in which case the Roth-Peranson algorithm of course cannot find a stable matching. But here magnitudes matter. Computational experiments showed that, under certain conditions that the market satisfies (not too great a proportion of couples and sufficiently short ROLs), as the market becomes large stable matchings exist with a high probability (Kojima et al. [2013]). The last lesson is that, *where preexisting theory is not decisive and in particular, where magnitudes matter, the designers made essential use of computational experiments.*

In fact, the NRMP has to date found stable matchings in each year of its operating. The algorithm 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]). The algorithm was implemented in 1998 and continues to be used to date.

[18] 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.

6 Meeting challenges from the philosophy of science

Recall the two challenges from the philosophy of science that we encountered earlier: GT models crucially rely on false assumptions, and they isolate mechanisms which are distorted by other mechanisms in the real world. The lessons from the previous section together with my account of GT models can now be applied to meet these challenges.

To be sure, in the simple matching model from the previous section, there are numerous assumptions which are likely, or even known to be, literally false. There are couples in the market, but the simple model assumes only individuals. Furthermore, the model makes assumptions concerning information that the agents privately hold, e.g. that they have transitive and complete preferences over agents on the other side of the market; that they are utility maximisers; that they know the "game" they are playing (the latter two assumptions were only implicitly made). These variables will likely not take those values in the target interaction. Do assumptions that are likely to be false, or even known to be false, not make it a "non-starter" to treat those assumptions as determining values in a causal graph?

I shall argue that they do not. My proposal is that, when modelling the interaction as a game, the values of the variables are tentatively fixed. In this process, things like simplicity, and deducibility, play a major role. For example, we started with a simple model without couples, and investigated what the policy goals could amount to there, as well as whether they can be implemented. So a simple model can direct attention to properties of potential importance for realising policy goals, as well as to game forms that implement them.

However, the conditions that define the simple model are not all (known to be) satisfied in the target interaction. This means that, when interpreting the graph of the model causally, some values are not known (as for variables that range over private information); or, it may be known that some values that variables take are not the values that they take in the target interaction (as for couples). The former case is an incomplete causal graph; in it, values of some nodes are simply not specified. And call the latter an imperfect graph, where values of some nodes are "false". Typically, the graphs of interest will be both imperfect and incomplete. These graphs can be read as suggesting causal hypotheses such as, "deferred acceptance algorithms cause agents to stick to the matchings". In causal hypotheses that an imperfect and/or incomplete graph suggests, the outcome may not follow deductively from its assumptions. For example, in the hypothesis above, nothing is said about agents' beliefs, preferences, etc., which are needed to deduce stability. In order to confirm such hypotheses that a given game form would indeed implement policy goals in the real world, various inference aids, such as experiments, can be used to establish inferences from the graph to the target interaction. Here, the lessons from the previous section kick in – as will be discussed in more depth below.

My proposal is akin to that of Anna Alexandrova, which is not surprising, as she is concerned with design economics, e.g. in Alexandrova [2006a,b, 2008]. She sees models as open formulae of the form, 'features F cause behaviours B in types of interactions x that have certain characteristics'. The proposition is open because it makes no existential claim that there exists a situation of type x. An open formula of interest could be, 'deferred acceptance algorithms cause agents to stick to matchings in markets where there is common knowledge of beliefs, preferences, etc. etc.'. Models as open formulae suggest causal hypotheses about the target interaction, namely that the target interaction is of a type where deferred acceptance algorithms cause agents to stick to the matchings. Importantly, not all the conditions that the model specifies need go into a causal hypothesis. It need not be assumed, for example, that there be common knowledge of beliefs in the NRMP's matching market. Without this assumption, the outcome may not deductively follow in the model. Empirical methods such as lab and natural experiments, randomised controlled trials, Mill's methods, and so on, are then needed to give evidence that the causal hypothesis does indeed hold.

I agree that the open-formula view does justice to the methodology of market design. It can be accommodated within my causal-graphs account; Alexandrova's open formulae correspond to imperfect and incomplete causal graphs. I believe that there is additional value to my graphical approach: graphs help uncover the causal structure that game theory assumes social interactions to have and which remains obscure when stated as open formulae. Unlike open formulae, graphs allow to show not only *that*, but *how* different game forms ('features F') and characteristics, including agents' preferences, beliefs, etc., affect model outcomes. Such model-interventions reveal counterfactual information that is vital for design purposes because it points at just what real-world interventions may bring about desirable outcomes.

It remains to consider inference aids more thoroughly. First, designers engage in experiments. Experiments are used for two things in this context: to fix values, and to give evidence that in a similar interaction to the one in question, unknown or false values do not disturb the causal mechanism much, which the model picks out. In our case study, natural experiments provided confidence that stable algorithms do in the real world achieve the policy goal that agents stick to the matchings. The logic seems to be something like the following: there are different social interactions that share some key features with the market of interest and which are likely subject to the same or similar causal mechanism. If there are variations between the different cases, then, if a model successfully accounts for the variations, it could be expected to predict the behaviour of other interactions that share key features with the investigated interactions. For example, there could be similar markets some of which produce reliable outcomes and some of which unravel. A model that accounts for the different outcomes, e.g. by pointing at a feature that all the reliable markets possess but the unravelling ones do not, is likely also a good model of a market that shares key features with the known markets. The other markets serve as natural experiments for the market at hand. Further, laboratory experiments are used to confirm that the difference between the markets is indeed the feature of interest that the model isolates.

What experiments do in this case is they give evidence that given values can be treated as approximations, and that the mechanisms which the model picks out is indeed structurally similar to the mechanism in the target interaction. They allow the inference that in markets with similar agents this assumption doesn't cause problems. In section 3, I emphasised that game theory is concerned with what types of social interactions result in what types of outcomes, because it is regularities that define types, which make such inductive inferences possible.

Designers also use the models themselves to establish inferences from the model to the target interaction. Some versions of such aids have been noted in the philosophy of science. To conclude, I shall discuss two such aids and show that designers use them where they are available; but also that they are not usually readily available. First, consider what Cartwright [1989] calls "concretization", and Hausman [1994], "de-idealization". I take these to be roughly equivalent; they amount to substituting unrealistic values of variables with more realistic ones, while preserving the outcome of the model. We encountered de-idealization when we considered too simple a model for the market for medical residents because couples are present in the market but not in the model. The model misrepresents the interrelated preferences of couples and instabilities resulting from them. The designers extended the model by adding couples to it. Now, this is not precisely de-idealizations because the outcome of the original model is not preserved: extending the model to accommodate couples shows that a simple deferred acceptance algorithm likely produces unstable matchings for couples, and that stable matchings may not even exist.

So, instead of de-idealization, we have failure thereof. This is not to say that the practice of replacing unrealistic assumptions with more realistic ones is not important. Extending the model did lead to a modification of the algorithm. But the case shows that often, replacing assumptions doesn't take the form of de-idealization. In particular, it draws attention to the importance of negative results. The impossibility result, "in markets with couples, there may not be stable matchings", motivated computational experiments on data to estimate magnitudes. So the model located problems, and suggested further computational analyses to resolve the issues. These, in turn, can prepare the ground for new theory. For example, the computational results suggested that there could be results proving that for large markets, it is unlikely that the set of stable matchings be empty. This intuition was indeed proven correct: about a decade later, Kojima et al. [2013] proved analytically the theorem that, if there are sufficiently small numbers of couples and the ROLs are short, as a market becomes sufficiently large the probability that a stable matching exists tends to certainty.

This method is not simply de-idealization. It rather draws a more complex picture, where substituting a true for a false assumption changes the outcome, which in turn leads to new theory through computational methods. Furthermore, de-idealization is not applicable for variables that take values in private information that is not known to the designer. This leads us to the last inference aid, which has already been mentioned in the description of how market designers treat preferences: robustness. This refers to the procedure of deriving the same or similar results from a wide class of models that differ in auxiliary assumptions, or from a very general model that can accommodate different auxiliary assumptions (e.g. Gibbard and Varian [1978], Sugden [2000], Kuorikoski and Lehtinen [2009]). Market designers make systematic use of robustness analysis when they design game forms which implement properties for variable populations and types of players. 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 LA, or the other way around. Nevertheless, it is assumed that the agents' preferences are complete and transitive. This is a milder assumption, even though it still is an assumption that may not hold for all agents in the market, and robustness only gets us so far.

Robustness analysis has struck some skepticism in the philosophy of science (e.g. Sugden [2000], Alexandrova [2006b]) on the grounds that it replaces false assumptions with assumptions that are not less false. For my purposes, the point is merely that inference aids such as robustness are used where they are available, but often they are not available, and so the causal interpretation of the graph requires more confirmation from elsewhere (a similar point is made in Alexandrova [2008]). Summing up, designers rely on inductive inferences from natural and laboratory experiments. In addition, they engage in analytic methods such as de-idealization and robustness, and they use computational experiments to estimate magnitudes that can't be inferred from the model. As Alvin Roth, the leading character of the case study, put it: "in the service of design, experimental and computational economics are natural complements to game theory" (Roth [2002]).

7 Conclusion

How do economic engineers harness GT models? I have argued that GT models assume that social interactions have a specific structure, which I depicted with the help of directed graphs. The contribution of game theory is that it allows to systematically track how model-interventions change model-outcomes. If the graph of a model can be interpreted as a causal graph of the target interaction, then model-interventions inform real world interventions.

However, many assumptions made in the model may distort the real world interaction that it is supposed to describe. In order to close the epistemic gap between the model and the real world, designers rely on inductive inferences from natural and laboratory experiments. In addition, they engage in analytic methods such as de-idealization and robustness, and computational methods.

My account of GT models makes precise an intuition that various philosophers of science have pointed at. According to Cartwright, models allow us to investigate isolated "capacities" Cartwright [1989, 2009]. Moreover, Mäki argues that models are needed to isolate causal mechanisms that are in effect in the real world but which are invisible-hand (e.g. Mäki [2009, 2011]). For both theorists, models isolate causal structures. One of the contributions of this paper is to illustrate a surprising new use of the directed-graph approach to uncover this specific causal structure for a class of GT models. I have shown how this philosophical tool can be lifted quite naturally out of standard game-theoretic practice.

Finally, my proposal is akin to those of Mary Morgan and Cartwright who have emphasised the fact that we do not just learn from models passively but by manipulating them (e.g. Morgan [2002]). In the words of Cartwright [2009], "[w]e change the models, experiment on them, see what results as assumptions are varied in relevant ways...we probe models as a means to understand *how* structure affects the outcomes." This intuition can be made precise. It amounts to model-interventions within my graphical representation. I believe that connecting insights from the two literatures of economic engineering and causal graphs has much potential to illuminate economic models, and graphs provide a particularly useful analysis from the perspective of policy-making.

This paper calls for various follow-up projects. Philosophers of science interested in economic engineering – mainly Alexandrova and Guala – have to date mainly studied the design of radio spectrum auctions, but not matching markets. What are the differences concerning the relation of models and experiments in these different subdisciplines of market design, and why do they arise? Furthermore, market designers seek to inform policy making; so in a sense it is no wonder that an interpretation of their models, which is explicitly causal and emphasises interventions, sits well with this practice. In what sense do models used in market design differ from models in other fields of economics, and can graphs shed light on models more generally?

References

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, 2006a.

Anna Alexandrova. *Connecting Models To The Real World: Game Theory in Action*. PhD thesis, 2006b.

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*. Oxford Handbooks Online, 2009.

Anna Alexandrova and Robert Northcott. Prisoner's Dilemma doesn't explain much. In Martin Peterson, editor, *The Prisoner's Dilemma*, chapter 4, pages 64–84. Cambridge University Press, Cambridge, 2015.

Robert Aumann and Adam Brandenburger. Epistemic Conditions for Nash Equilibrium. *Econometrica*, 63(5):1161–1180, 1995.

James M. Buchanan. Game theory, mathematics, and economics. *Journal of Economic Methodology*, 8(1):27–32, 2001.

Nancy Cartwright. *Nature's Capacities and Their Measurement*. Oxford University Press, Oxford, 1989.

Nancy Cartwright. *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press, Cambridge, 1999.

Nancy Cartwright. If no capacities then no credible worlds. but can models reveal capacities? *Erkenntnis*, 70(1):45–58, 2009.

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

Allan Gibbard and Hal Varian. Economic models. *Journal of Philosophy*, 75:664–677, 1978.

Francesco Guala. How to do things with experimental economics. In Donald MacKenzie, Fabian Muniesa, and Lucia Siu, editors, *Do Economists Make Markets? On the Performativity of Economics*, chapter 5, pages 128–162. Princeton University Press, Princeton, 2007.

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

Daniel M. Hausman. Paul samuelson as dr. frankenstein: When idealizations escape and run amuck. In Bert Hamminga and Neil de Marchi, editors, *Idealization in Economics*, pages 229–243. Poznan Studies in the Philosophy of the Sciences and the Humanities. Amsterdam: Rodopi, 1994.

Daniel M. Hausman. *Taking the Prisoner's Dilemma seriously: what can we learn from a trivial game?*, chapter 3, pages 54–63. Cambridge University Press, Cambridge, 2015.

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.

Shengwu Li. Ethics and market design. *Oxford Review of Economic Policy*, 33(4):705–720, 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.

Mary Morgan. Model experiments and models in experiments. In L. Magnani and N.J. Nersessian, editors, *Model-Based Reasoning: Science, Technology, Values*. Kluwer Academic/Plenum Publishers, New York, 2002.

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

John F. Nash. Two Person Cooperative Games. *Econometrica*, 21:128–140, 1953.

Martin J. Osborne and Ariel Rubinstein. *A course in game theory*. The MIT Press, Cambridge, Massachusetts, 1994.

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

Dani Rodrik. *Economics Rules.* W.W. Norton & Company, 2015.

Alvin E. Roth. The Economics of Matching: Stability and Incentives. *Mathematics of Operations Research*, 7(4):617–628, 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*, 36:277–288, 1985.

Alvin E. Roth. Game Theory as a Part of Empirical Economics. *Economic Journal*, 101:107–114, 1991.

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*, 289:909–912, 2003.

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*, chapter 1. Oxford University Press, Oxford, 2013.

Alvin E. Roth. *Who Gets What - And Why: The Hidden World of Matchmaking and Market Design.* Eamon Dolan / Houghton Mifflin Harcourt, 2015.

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: 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.

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

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

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

Roberto Serrano. Fifty Years of the Nash Program, 1953-2003. *Investigaciones Economicas*, XXIX(2):219–258, 2005.

Peter Spiegler. *Behind the model: a constructive critique of economic modelling.* Cambridge University Press, Cambridge, 2015.

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

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

John Sutton. *Marshall's Tendencies: What can economists know?* MIT Press, Cambridge, Massachusetts, 2002.

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

James Woodward. *Making Things Happen – A Theory of Causal Explanation.* Oxford University Press, Oxford, 2003.

Petri Ylikoski and Emrah Aydinonat. Understanding with theoretical models. *Journal of Economic Methodology*, 21(1):19–36, 2014.