The virtual reality of the invisible hand

DAYER-TIEFFENBACH, Emma

Abstract

Self-centred based explanations such as invisible-hand accounts look like armchair constructions with no relevance to the real world. Whether and how they nonetheless provide an insight into social reality is a puzzling matter. Philip Pettit’s idea of self-interest virtually bearing on choices offers the prospect of a solution. In order to assess the latter we first distinguish between three variants of invisible-hand explanations, namely: a normative, an historical and a theoretical one. We then show that, while the model of virtual self-interest is a helpful gloss on each variant, it may not convincingly succeed, pace Pettit, in reconciling the economic mind with the common mind.

Reference


DOI : 10.1177/0539018415608966
The virtual reality of the invisible hand

Emma Tieffenbach
University of Geneva, Switzerland

Abstract
Self-centred based explanations such as invisible-hand accounts look like armchair constructions with no relevance to the real world. Whether and how they nonetheless provide an insight into social reality is a puzzling matter. Philip Pettit’s idea of self-interest virtually bearing on choices offers the prospect of a solution. In order to assess the latter we first distinguish between three variants of invisible-hand explanations, namely: a normative, an historical and a theoretical one. We then show that, while the model of virtual self-interest is a helpful gloss on each variant, it may not convincingly succeed, pace Pettit, in reconciling the economic mind with the common mind.

Keywords
Carl Menger, invisible-hand explanation, Philip Pettit, resilience, self-interest

Résumé
Les explications qui font l’hypothèse de motivations exclusivement intéressées de la part des individus, comme les explications dites ‘par la main invisible’, paraissent dénuées de pertinence eu égard au monde social réel. Dans quelle mesure ces explications éclairent tout de même la réalité est une question épineuse. L’idée défendue par Philip Pettit selon laquelle l’intérêt personnel influence virtuellement les choix se présente comme une solution. Pour l’évaluer, nous distinguons d’abord trois variantes des explications par la main invisible en fonction des buts respectivement normatifs, historiques et théoriques que poursuivent ceux qui les avancent. Nous montrons ensuite que le modèle de la réalité virtuelle de l’intérêt personnel permet d’élucider chacune de ces variantes sans toutefois parvenir de façon convaincante à réconcilier la psychologie économique et celle du sens commun.

Corresponding author:
Emma Tieffenbach, Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland.
Email: Emma.tieffenbach@unige.ch
Invisible-hand explanations look very much like armchair constructions, with no relevance to actual historical events. The story they tell shows how agents can contribute to the establishment of some presumably desirable social outcome while being moved by no other purpose than that of satisfying their self-centred preferences. Legislators and binding agreements play no role in these highly parsimonious explanations, which also further economize on knowledge when they assume that agents are not even aware of the institutional patterns that their choices ultimately bring about. Surely, the various disinterested motives that are left aside in these stories play a role in real-life deliberations. Yet it is the very purpose of these explanations to ostensibly dispense with these types of deliberations, however more plausible they admittedly are, and, even more crucially, to have us count them as useless cogwheels of the social machine.

Apart from the undeniable aesthetic achievement of explaining a lot by a little, one may legitimately ask whether dispensing, as invisible-hand explanations typically do, with more realistic motivational assumptions provides any insight into the social reality. This is obviously a puzzling issue and, unsurprisingly, also a dividing topic, those who take a stance on the matter falling somewhere on a continuum between a sceptical assessment (cf. Bunge, 1997; Elster, 1989; Rosenberg, 1995), on the one hand, and a rather enthusiastic, if not programmatic, appraisal (cf. Aydintonat, 2008; Nozick, 1974; Ullmann-Margalit, 1978; Vanberg, 1994), on the other.

In a series of chapters and articles, Pettit provides the elements of what seems to be an ingenious way of approaching the problem (Pettit, 1996, 2000, 2001, 2007; see also Brennan & Pettit, 2005). His suggestion is to take the self-centred assumption as having a purely virtual presence in ordinary human deliberation, and explanations that are based on this assumption as being about the ‘resilience’ of social patterns. Pettit thought that not only rational choice explanations but also the equally controversial classes of functional explanations and game theoretical approaches could be re-interpreted in the same manner (Pettit, 2000). The possibility that invisible-hand explanations may be similarly rescued from the problem of the empty ‘black box’, i.e. assuming a mind that flies in the face of common sense and plausibility (Pettit, 2000), deserves consideration. The idea seems plausible enough given the central role that self-centred motivations are assumed to play within these explanations. Moreover that Pettit’s revised rational choice framework could be appealed to for this particular purpose is a possibility at which Pettit hints (Pettit, 2000: 37). Providing the relevant awaited development is what this article aims to do.

The discussion will proceed as follows. The first section provides a workable definition of invisible-hand stories. After listing a few of their desiderata in our first section, three variants of invisible-hand explanations are distinguished. Taking Menger’s account of the emergence of the money system as an example, we show that one and the same invisible-hand explanation may be approached as either describing a welcome possibility (this is the normative application), the historical path by which some particular social
pattern arose (this is the historical use) or, still alternatively, being a ‘process-defective’ account (Nozick, 1974: 8) whose unclear explanatory value has elicited the conflicting views previously outlined. After outlining Pettit’s revised rational choice framework, we show why it is a helpful elucidation of each variant of invisible-hand explanations. We finally clarify the sense in which self-centred motives have a virtual presence in agents’ mind. While Pettit’s theory was prompted by the idea that relying solely or predominantly on selfish motives flies in the face of common sense, we show that to ascribe to selfish motives a virtual role may be just as un-commonsensical.

**Invisible-hand stories: A few desiderata**

Invisible-hand explanations come in the form of narratives that are easily recognizable by virtue of a few common features. In order to spell out the latter, Menger’s account of money (2007 [1871], chap. 8, 1892) will be the template. Menger describes a process by which the barter system is replaced by one in which a good is generally accepted as a medium of exchange. When agents barter their goods, he first observes, they act in light of their immediate needs: ‘Each man intends to get by way of exchange just such goods as he directly needs, and to reject those of which he has no need at all, or with which he is already sufficiently provided’ (Menger, 1892: 241). Yet the barter system turns out to be an inconvenient way of acquiring goods. This is because, for an exchange to happen, it is necessary to meet someone who has what one needs and who needs what one has. This double coincidence of wants is unlikely to take place as often as it ideally should (1892: 258). As a consequence a lot of valuable time is often required to carry out a trade, and ‘the number of bargains actually concluded must lie within very narrow limits’ (1892: 258). Proceeding with indirect exchanges is the solution: instead of accepting the good they presently desire to have, it is in agents’ interest to provisionally accept another good with a view to exchanging it later for the good they are looking for. This intermediate good must be well chosen, however. It will play its function of reducing the trading costs only if it is itself a frequently traded good. In Menger’s terms, the intermediate good must have a high level of ‘saleableness’ (1892: 248). This way it is more likely to be accepted by those who happen to own the good that one wants. Possessing a few goods that present these features is thus an advantage. Now each time a good is chosen as an intermediate good, its saleability rises. But its saleability was already the reason why it was initially chosen as an intermediate good. This is therefore a self-reinforcing process. The more a given good is chosen as an intermediate good, the more its saleability increases, and the more it will henceforth be chosen as an intermediate good. Ultimately agents accept a small number of goods (only one in some cases) as the dominant intermediate good(s). Following Menger, one can conclude that money has arisen when many (if not all) converge on a few (if not only one) intermediate goods.

Menger’s explanation is certainly at odds with a more intuitive way of explaining the institution of money. It noticeably dispenses with the factors that are expected to play a role in the bringing about of its *explanandum*. In particular it does not refer to a central authority which guides agents through their choices. Nor do agents have the opportunity of meeting and discussing the matter. No agreement (no ‘common will’, as Menger says) fixing a medium of exchange ever took place between them. In fact agents do not even
have in mind the overall consequence of what they individually do. No one needs to grasp the final stage of the process by which money is brought about in order to be active in its establishment. Agents only need to pursue the prosaic goal of increasing the number of their possessions, and to be endowed with the capacity to perceive the benefit of using a good as a medium of exchange for that purpose. Money, on this view, is ‘the unintended result of innumerable efforts of economic subjects pursuing individual interests’ (Menger, 1985 [1883]: 147, 158). Note that the assumption that agents are self-interested is intended to rule out an agreement or a central authority. To clarify this point one may distinguish between two ways the self can play a role in the shaping of goals. Agents may be self-determined and/or have a self-centred goal. Menger’s explanation, as we shall see, is committed to both senses.

First, agents are self-determined when, pursuing their own goal, they do not execute someone else’s plan. The requirement is one of autonomy and it is one that is clearly not fulfilled when agents realize someone else’s order. Menger’s explanation clearly conforms to the self-determination principle when it describes agents as acting on their own initiative.

Second, agents have self-centred goals when they act on the basis of their own goals rather than on the consideration of other’s. While the self-centred assumption allows these goals to be benevolent or altruistic ones, there also are goals that agents will describe as their own. The requirement is not fulfilled when agents register the goals of others and commit themselves to further them (cf. Sen, 2002). Nor is the principle of self-centred goal satisfied when agents see what they individually do as their participatory actions in the realization of a shared goal (Anderson, 2000; Schmid, 2005). That agents do not incorporate within their choice another’s choice, in either of these two senses, is also true of agents involved in the emergence of money, in Menger’s description of the process.

Did most money systems emerge this way? The point can be argued (cf. Graeber, 2011). But our interest lies in the nature of the explanation that is advanced, not in its accuracy or questionable validity (cf. Bridel, 2014, on this latter point). Following Nozick (1974; see also Aydinonat, 2008; Haller, 2000; Vanberg 1994), we take Menger’s account as a paradigm of invisible-hand explanations, therefore exhibiting most of the ingredients that are distinctive of this class of explanations. When used for such illustrative purpose, the following features may in particular be retrieved from it. First of all, in order to explain how the money system arose, Menger does not merely point to the deficiencies of the barter system, thus leaving one to guess how precisely the latter finally was replaced by the former. Rather, the mechanism he proposes details the various steps between these two stages. These steps are, in turn, spelled out in terms of the various strategic choices, beliefs and expectations leading agents to behave in the way that is required for money to arise. In line with methodological individualism, Menger provides a fine-grained explanation that searches for the micro-foundations of social regularities. Yet invisible-hand explanations are individualistic in still another sense: they assume that agents do not act jointly, in any of the various senses that are offered of this debated notion. In contrast to various recent accounts of money (see Searle, 1995, 2010; Tuomela, 2003), Menger’s explanation conspicuously dispenses with any form of collective intentionality, relying only on individual intentions in its description of the process resulting
in the creation of the money system (see Smit et al., 2011; Tieffenbach, 2010; Turner 1995: 225). What Menger shows is that agents’ preference will spontaneously converge on the same commodity in the absence of any coordinative effort on their part.

While an individualistic explanation may consistently refer to its explanandum as what agents deliberately intended to produce, an invisible-hand explanation describes it as an unintended consequence. Money is indeed the effect of everyone’s choice without anyone having intended it. The unintendedness of the outcome to be explained is sometimes said to confer invisible-hand explanations their particular explanatory import (Nozick, 1974; Ullmann-Margalit, 1978; Williams, 2002). As Nozick explains, these explanations ‘don’t explain complicated patterns by including the full-blown pattern-notions as objects of people’s desires or beliefs’ (Nozick, 1974: 14). In a ‘straightforward explanation’, by contrast, part of the explanandum appears in the explanans as what agents intend to bring about, therefore presupposing some sort of knowledge of that explanandum. Invisible-hand explanations make no presupposition of that sort because they cannot, as a constraint, represent the outcome to be explained as the content of agents’ intentions. Their explanans will only include what agents intend, thus avoiding any reference to the explanandum. For this reason they can further be characterized as perfectly non-circular explanations or, as Nozick calls them, ‘fundamental explanations’ (Nozick 1974: 8-9).

Now the idea that agents are not aware of what their actions ultimately amount to is also a feature of functionalist explanations (Elster, 1986). Functionalist explanations account for a social pattern (e.g. a rain-making dance) by pointing to its desirable, albeit unintended, consequence (e.g. greater social cohesion). They sometimes are treated as particular cases of invisible-hand explanations (Nozick, 1974: 19; Rosenberg, 1995; Ullmann-Margalit, 1978). However, they really are conceptually distinct from, rather than a variant of, the latter. Both explanations refer to unintended consequences, but the latter occupy two different places within in each of them. In a functionalist explanation, the outcome to be explained brings about unintended consequences, which retroactively perform the explanatory work (i.e. the explanans). In an invisible-hand explanation, the unintended consequence is the outcome to be explained (i.e. the explanandum).

To sum up, explanations in the style of the invisible-hand are at best parsimonious, individualistic, fine-grained and non-circular accounts. These features help single out a variety of explanations without being strictly discriminatory. They are not essential ingredients, but mere desiderata, i.e. requirements that are desirable and plausible, without being strictly disqualifying.

A multi-faceted notion

Invisible-hand explanations may be offered for some other purpose than that of explaining. They may, in particular, be intended as describing a commendable way of obtaining a certain social outcome. Menger’s explanation, for example, may be approached as pointing to the futility of a decree regarding what should be used as money. The idea that, as Menger says, ‘money is not an invention of the state. It is not the product of a legislative act’ (Menger, 2007 [1871]: 261) may thus be taken both as a descriptive and as a normative statement. Because Menger’s explanation ontologically dispenses with a central authority, it naturally questions the practical usefulness of the latter. In particular,
Menger seems to invite us to believe, by way of his explanation, that agents need not be externally obliged to observe the rules of an institution which, following only their material interest, they have a reason to conform to, even in the absence of such external obligation. An enactment making a certain good the official medium of exchange would, at most, add stability to an institution that already works by itself. The state is, however, not entirely deprived of a role in the full establishment of the money system. Its import is that of making ‘legally binding’ (Menger, 2007 [1871]: 262) the use of the unique good it has authorized, converting it into a ‘universal substitute in exchange’ ([2007 [1871]: 262). As Menger says, while the state ‘is not responsible for the existence of the money-character of the good, it is [nonetheless] responsible for a significant improvement of its money-character’ (Menger, 2007 [1871]: 262). There lies the merit of legislators: in helping the money system be fully realized by ensuring that only one good is ascribed the unique function of money by everyone.

Presumably the freedom and autonomy of agents involved in an invisible-hand process make the latter an attractive path. The normative dimension of Menger’s explanation may not, however, be featured in all explanations of the same type. Yet more or less implicit in invisible-hand accounts is the idea they are about outcomes that could not have materialized were it not for the invisible-hand process, or that they are, at least, best produced by means of it. While the state can in principle intend to create money by simply printing fiat money and making it legal tender, Menger’s explanation shows that the latter will be money only if each agent has an independent, self-interested reason for individually using it as such. Even more crucially it shows that such a reason prevails regardless of any decree. When they have such a policy orientation, invisible-hand explanations are works of institutional design.\footnote{5}

A second, broadly conceived, use of the notion is exemplified by the cases where the invisible hand is invoked with a plain explanatory intention. This time the idea is not, or not only, to recommend something, but to account for, to describe or to report. One further distinction needs to be made, within this broad class of explanatory applications, between an historical and a theoretical variant. In his methodological writings, Menger offers some ways of approaching the difference between these two enterprises (Menger, 1985 [1883]). He says that a social structure, like an institution, can either be ‘historically’ or ‘theoretically’ understood, and that these ‘two great classes of scientific knowledge’ (1985 [1885]: 35) correspond to two different ‘points of view’ (1985 [1883]: 35). One can adopt either, he also argues, depending on whether one is interested in examining the ‘individual’ aspects of phenomena or, on the contrary, aiming to grasp their ‘general’ features. The two approaches are thus not mutually exclusive and even complement each other (see Haller, 2004).

Menger also says that the individual aspects of a social phenomenon are known by ‘investigating its individual process of development’ (Menger, 1985 [1885]: 43). This requires us to pay attention to ‘the concrete relationships under which it has developed and, indeed, has become what it is, in its special quality’ (1985 [1885]: 43). Menger describes the task of the historical sciences as providing an understanding of ‘concrete phenomena, located in space and time’ (1985 [1885]: 35, 37). Historical sciences are about the ‘particular’, the ‘definite’, the ‘individual’ or, as he also says, the ‘specific’, the ‘immediately experienced’ (1985: 56). Their scope is social phenomena taken as tokens,
that is, as non-repeatable particulars. For example, we can investigate how the money system emerged in some particular prisoner of war camp during the Second World War. This investigation will have to depict the particular process by which the prisoners managed, without any consultation meeting, and, of course, without any external coordinator, to organize amongst themselves in a way that finally led them to attribute to the very same commodity (e.g. bread, cigarettes, etc.) the function of serving as a medium of exchange.

The use of a commodity as a medium of exchange occurred many times. Other similar, special cases of commodity goods come to mind, such as the use of cowry shells by the Ojibway, the use of packs of Marlboros in Moscow in 1990, squirrel furs in medieval Finland, bird scalps among Californian tribes in the 19th century, or, to take Menger’s list, the use of coconuts, ice-cubes, cattle, skins or tea bags in other regions. From this point of view, the use of bread in prison camps still presents itself as a token, but as a token of a type, and one’s attention now switches to the latter perspective. As Menger says, ‘definite phenomena are repeated now with greater exactitude, now with lesser, and recur in the variation of things. We call these empirical forms types’ (1985 [1885]: 36). Types are the recurrent aspects of things. It is by virtue of seeing that not only cowry shells but also many other commodities were repeatedly purchased for the same purpose, i.e. serving as media of exchange, that money as a type can be discovered.

Once the bread and its users are recognized as instances of types, one switches from the historical perspective to a theoretical one. The specific task of the theoretical approach is to achieve an understanding of the money system that will not, or not directly, pertain to its diverse historical instantiations. The theoretical approach requires the ability to abstract some non-causal, or, as Menger says, the ‘necessary’, (1985 [1885]: 60) connections between types (between money, economic agents, exchanges, goods, prices, etc.) from the particular development of the concrete social phenomena that are examined. I take Menger to say that theoretical explanations are metaphysical explanations in which the explanans and explanandum are connected, not through some kind of causal mechanism, but through some ontological link by which the essential, necessary connections between types are elucidated (see Mäki, 1990a, 1990b, 1997). Menger expected a theoretical explanation to provide an understanding of the ‘nature’ or ‘essence’ (Menger, 1985 [1885]: 125) of social phenomena. Invisible-hand explanations may be, and in fact often are, illustrative of this theoretical variant. Menger’s account of the origin of money may be considered as an illustration (like Hume’s explanation of the emergence of property rights and Nozick’s story of the rise of the minimal state).

In sum, various types of invisible-hand explanations can be distinguished according to the purposes they serve. Note that it is easy to be fooled by the descriptive form the normative variant of the notion may display. One may conversely fail to discern that a seemingly historical explanation is, in fact, an embodied theoretical explanation. Invisible-hand explanations can prove deceptive.

**The virtual reality of self-interest**

A question that remains to be addressed is whether and how the self-interest assumption that is featured in all three variants of invisible-hand explanations in the previous sections serves the various normative and explanatory purposes at stake in each
We shall turn to Pettit’s idea of self-interest virtually bearing on agents’ mind as a way of dealing with this issue. Pettit first observes two apparently non-reconcilable facts about the self-interest assumption. One is its empirical inaccuracy: a broad range of human interaction shows that human beings are moved by considerations that are far from being exclusively egocentric. The postulated self-interested mind ‘is not the mind that people manifest in most social settings and, in particular, that is not the mind that common sense articulates’ (Pettit, 2001: 308). As Pettit argues, the latter would rather have loyalty, fair play, kindness, friendship, politeness, honesty, merits and straight talking as the kind of deliberative themes that bear on our action. The items in this list of motives are quite disparate. Yet, they all have in common that they are ‘concerns that dramatically transcend the boundaries of the self’ (Pettit, 2001: 317). The other fact is the presence of the self-interest assumption in most economic explanations. Pettit argues that the discrepancy between what common sense establishes and what economic explanation presupposes raises some doubt about the ability of the latter to be of explanatory value as regards real-world situations:

The mind postulated in rational-choice theory is that of a relatively self-regarding creature. But the mind that people display towards one another in most social settings, the mind that is articulated in common conceptions of how ordinary folk are moved, is saturated with concerns that dramatically transcend the boundaries of the self. So how can we invoke the workings of the economic mind to explain behaviour, when the ‘black box’ at the origin of behaviour does not apparently contain an economic mind? (Pettit, 2000: 40)

Pettit’s model is meant to protect economic explanations from the defect of being of the ‘empty black-box’ kind, that is, from being causal explanations that postulate a mind that mismatches the one with which we are daily acquainted. Pettit’s solution requires that we should not take these explanations at face value. In particular, self-interest, according to Pettit, does not have to be the actual cause of human behaviour because it may instead have a virtual presence in people’s minds. Pettit’s version of rational choice theory introduces a significant twist to the conventional view by making one of its alleged central claims, namely, the self-interest assumption, a description of how agents behave virtually rather than how they actually behave. If self-regarding considerations have a virtual presence, they are not the causes of actual behaviour. Rather they are ‘virtual causes’, ‘standby causes’ (Pettit, 2000: 44, 45, 46), ‘potential causes’ (2000: 44) or, as Pettit also calls them, ‘causes of a would-be variety’ (2000: 44).

We are asked in particular to re-read their factual statements as counter-factual ones, thus translating the various claims about what agents do as referring to what they would do if they ceased to act in the way they actually behave. Such a modal switch, Pettit says, will transform the self-regard assumption into one that no longer conflicts with our everyday experience (2000: 43). Self-regard, virtually construed, affects the way people behave only in circumstances when their personal advantage is compromised. At that point, loyalty, kindness, politeness and honesty give way to self-centred thoughts. Pettit uses a number of lively metaphors that help visualize the way self-regard bears virtually on actions. One of the most telling is about alarm bells ringing:
There are certain alarm bells that make [agents] take thought to their own interests. People may proceed under more or less automatic, cultural pilot in most cases but at any point where a decision is liable to cost them dearly in self-regarding terms, the alarm bells will tend to ring and prompt them to consider personal advantage; and heeding considerations of personal advantage will lead people, generally if not invariably, to act so as to secure that advantage: they are disposed to do the relatively more self-regarding thing. (Pettit, 2000: 240)

To paraphrase: selfless preferences have a bearing on what people do only insofar as their own self-regarding interests remain protected. It is when the latter are jeopardized that people start calculating the personal losses and benefits occasioned by their behaviour. ‘Self interest will kick in to modify action’, Pettit claims, ‘in cases where the actor comes to realize that the cost of failing to take interests explicitly into account has become too great’ (Brennan & Pettit, 2005: 277). There is a threshold effect that alters the framework for our decision-making deliberations. We stop caring about other-regarding goods if and when we come back down to the threshold where self-interested goods start having marginal utility again. And we do so either because we lost some goods, faced the prospect of losing some or got greedier.

The model accommodates subjective variations. While some agents will sense that their personal interest is jeopardized as soon as the slightest loss occurs, others will wait for a large amount of self-sacrifice before reaching the same conclusion. Pettit’s theory provides room both for the robust altruist and for their less kind-hearted counterpart. The fact that they do not take the same view about when their self-interest is threatened does not prevent the latter from playing its virtual role. In sum, all that is needed for self-interest to be virtually recognized and acted upon is that it is subjectively felt to be threatened. The cases where agents seem to deliberate in a way that compromises their self-interest only reflect the unusually high cost at which they still are willing to secure it.

Let us grant that self-interest can play a virtually deliberative role. The next pressing question is: What can assuming things in this manner possibly explain? In order to address it, Pettit distinguishes between three types of *explananda*. We can first be interested in explaining the emergence (or the appearance, establishment, institution, occurrence, bringing about) of a social pattern. Second, we may be interested in explaining its continuation (or its maintenance, duration, persistence). Or we could be concerned to explain its resilience (or its robustness, stability). To explain the emergence and maintenance of a social pattern, one will have to trace its actual causal story. To explain the resilience, one will rather have to point to its virtual cause. The latter will show ‘why it is such that for a variety of possible scenarios, possible ways the world may go, it remains a fixture’ (Pettit, 2007: 83). Pettit’s argument can be summarized by way of the following premises and conclusion:

1. Economic explanations represent agents as being exclusively self-centred.
2. The self-centred assumption flies in the face of common sense, and therefore economic explanations raise the problem of the empty black box.
3. The self-centred assumption offers a more plausible view about the way the mind works once it is modally qualified as being a virtual cause of our behaviour.
4. The virtual cause of an outcome explains its resilience.
Some steps of this argument can be disputed. Premise 1 is a factual statement about the place of the self-interest assumption within economic theory. Many will reply that the utility-maximizing principle predicts only that people do what they subjectively prefer doing (Hausman & McPherson, 2006). The possibility of altruistic behaviours is left open if that happens to be what agents prefer doing. Utility, it is argued, will be treated as an empty placeholder, allowing for any kind of mental or physical states to be the object of maximization. People who rank the well-being of others above their own are no less utility-maximizers than those who have an exclusive concern for their own well-being. Anticipating this objection, Pettit argues that economists are implicitly committed to the self-interest assumption in their theoretical tools (e.g. the Paretian assumption and the downward-sloping demand curve), despite their denial in this regard (Brenan & Pettit, 2005; Pettit, 2000). Premise 2 is the rather uncontroversial view about the lack of empirical adequacy of the self-interested assumption. Whether Pettit’s notion of self-interest offers a more plausible view is discussed in the last section of this article, where we take a closer look at Premise 3. Before, we need to show why invisible-hand explanations are *prima facie* eligible for Pettit’s revised rational choice framework.

**A re-appraisal of the three variants of invisible-hand explanations**

We have seen how self-interest can be construed as having a virtual present in agents’ mind. In this section, we show that the idea is very consistent with, and in fact may help elucidate, each of the three variants of invisible-hand explanations that were previously distinguished.

When they serve a normative goal, invisible-hand accounts indicate how some presumably desirable outcome ought to be obtained. They thus point to a desirable possibility. At first sight, the assumption that self-regard can virtually bear on agents’ choice does not seem to fit this normative agenda. Indeed, the possibility that agents may act self-interestedly is, to say the least, not a particularly attractive one. There is a sense in which the possibility may, however, be found appealing. Suppose, as Pettit does, that self-interest could sustain the very same behaviour that agents currently engage in for selfless reasons. What makes the former case a welcome prospect is that it is one we can rely on in circumstances where agents happen to lack the ability to act in a more commendable fashion. It represents a valuable possibility insofar as it offers us the possibility of not having to rely exclusively on more demanding character traits and motivations. It depicts a process by which an outcome could, in the worst-case scenario, be obtained should agents fall short of more appealing motivations. It also offers the prospect of designing institutions that are not too demanding with respect to what agents are able to do. Echoing the knave principles advocated by Hume, Pettit says that the assumption is in this case devised with the objective of finding those institutions that are ‘likely to prove more durable in face of shifting fashions and uncertain levels of human commitment’ (Pettit, 2007: 83). The self-interest assumption virtually understood thereby fits the normative concern with which the invisible hand is sometimes invoked.
Let us now consider the explanatory purpose to which the invisible hand is alternatively dedicated. As suggested earlier, such an explanatory purpose splits into two kinds, namely, the historical and the theoretical variants, each of which will be considered in turn.

In order to see how virtual self-interest serves the historical variant, we shall first see why a causal explanation of any particular event may take the unexpected form of a conjecture. Conjectures are arguably useful when historical evidence is insufficient to reconstruct accurately how things happened. This is the ground on which Dugald Stewart’s idea of conjectural histories is justified (Stewart, 1858: 34): a second-best explanatory strategy. The realm of conjecture is what the historian resorts to in the absence of evidence about how things actually happened. Venturing into conjecture, Stewart argues, is better in all cases than invoking a miracle. Because of their famous lineage in the Scottish Enlightenment, invisible-hand explanations are sometimes regarded as illustrative of Stewart’s idea of conjectural histories (see Aydinonat, 2008; Barry, 1982: 48; Hopfl, 1978; Keller, 1994). When approached as conjectures, invisible-hand explanations represent ways of bridging the gap in the documentary record. They track reality, and fail to do so only when data are lacking.

The necessity to venture into a conjecture when lacking evidence is a faute de mieux. What remains to be explained, however, is why conjectures have to be about the self-interested nature of agents’ motivations. The following rationale may be given. One reason for assuming self-centred motives is related to the deceptive nature of disinterested behaviours. We often do not know for sure whether a certain apparently other-regarding behaviour is based on a genuinely disinterested motivation. Absent any direct evidence, caution may require one to assume that selfishness is at play (Elster, 2009). Another reason to assume self-regard is that it is the only inclination that could be universal. As Elster says, ‘the goal of the altruist is to provide others with an occasion for selfish pleasures … If nobody had first-order, selfish pleasures, nobody could have higher-order, altruistic motives either’ (Elster, 1983: 45). In other words, while the altruists need the selfish to be who they are, the reverse is not true. These two considerations may not, by themselves, be sufficient to fully justify the widespread use of the self-regard assumption. They may, however, speak in favour of such a use in the particular situation where the effectiveness of disinterested motives is not supported by evidence. Note, however, that, in this case, the success of the conjectural account that is provided will depend on the way it succeeds, even if only by chance, in tracking such an historical reality. A good conjecture is one that accidentally hits the truth.

Let us finally consider the theoretical mould in which an invisible-hand explanation may still alternatively be cast. This time, getting the facts right adds little, if any value at all, to their explanatory power. Nozick makes this point straightforwardly. The invisible-hand explanations he provides of the emergence of the minimal state, he says, will ‘serve our explanatory purposes … even if no actual state ever arose that way’ (Nozick, 1974: 7). In the same vein, he also says that ‘we learn much by seeing how the state could have arisen … even if it didn’t arise that way’ (1974: 9). His explanation is that of a ‘process-defective account’ (1974: 9), and that is certainly not a defect capable of being rectified by way of conjectures in the hope that one luckily hits on the truth and provides the process-accurate account that, in the present state of historical knowledge,
is unavailable. Surprisingly, Nozick does not specify what we precisely learn by the sort of conjecture or process-defective account he ventures into. A broad answer to that question is proposed by Ullmann-Margalit when she notes how an historically false invisible-hand explanation may be helpful in highlighting the ‘self-reinforcing’ nature of its outcome:

Even if the invisible-hand explanation turns out not to be the correct account of how the thing emerged, it may still not be devoid of validity with regard to the question of how (and why) it is maintained. … The availability … of a cogent invisible-hand story of how the pattern in question could have arisen – given the specific circumstances, some common-sense assumptions concerning the drives of the individuals concerned, and the normal course of events – may, I believe, contribute to our understanding of the inherently self-reinforcing nature of this pattern and hence of its being successful and lasting. (Ullmann-Margalit, 1978: 275; my emphasis)

Note that when Ullmann-Margalit speaks about the way a pattern is ‘maintained’ or ‘lasting’, she is merely referring to the possibility of its persistence throughout a certain period of time. But when she talks about the ‘inherently self-sustaining’, ‘self-reinforcing’ or ‘successful’ feature of such a pattern, she is distinctively referring to its stability, that is, to its capacity to remain in existence in a wide range of circumstances. A lasting social pattern might be very fragile and thus inclined to collapse as soon as it encounters some perturbing factors. It might be that luck or the mere fact that no destabilizing event ever occurred accounts for its maintenance. Consider, as an example, a social pattern that is obtained in virtue of the mechanism of preference falsifications (cf. Kuran, 1995). This pattern lasts because agents do not dare to publicly express what they intimately prefer, and they hide their preference because they mistakenly believe them to be held by no one but themselves. Such a pattern may well last for decades (the millenary cast system, Kuran suggests, should be explained in this manner) while not being stable. Let some agents stop falsifying their preference and make the real content of the latter be made public and the system will collapse. An ‘inherently self-sustaining’ institution is, by contrast, one that succeeds in overcoming unsettling events or factors. It is, in other terms, a stable or, as Pettit says, a ‘resilient’ social phenomenon. Distinguishing between the continuance, on the one hand, and the robustness of a social pattern, on the other hand, is of obvious conceptual interest. More importantly, this distinction also suggests where the explanatory value of the theoretical variant of invisible-hand explanations resides. It points to the possibility that the latter distinctively accounts for the stability (or ‘self-reinforcing’ nature) of the outcome, leaving the question of their emergence (or, for that matter, of their continuance and, we may also add, of their disappearance) to those who take a historical point of view on the matter.

Menger’s explanation of the money system could be re-interpreted in accordance with Pettit’s notion of resilient explanations. The idea is that, even if Menger is wrong when he describes money as being neither the product of a common will nor the effect of a legislative decision (even if, in other words, it turns out that agents use certain things as media of exchange because they are ordered to do so or, alternatively, because they manage to deliberately coordinate themselves to this effect), it does not imply that self-centred considerations do not play any role in the resilience of the actual money system. What
Menger shows is that, if agents stopped being moved by the sort of selfless considerations (that is, by a decree or by a previous binding agreement) that currently move them, they would have an additional, individual reason to keep the money system. Once Menger’s explanation is re-interpreted in this manner, however, what it explains should also be accordingly re-appraised. While being of no help in understanding why the money system emerged (nor, for that matter, why it continued), Menger’s model may be of interest to predict why such a system, once instituted, is likely to remain as it is.

We have shown in this section how Pettit’s idea of economic explanations coheres with some of the purposes for which the invisible hand is invoked. Assuming the virtual reality of self-regard turns out to be a particularly useful methodological device in the circumstances when:

1. designing institutions, we would like to leave room for the knaves among us;
2. explaining the emergence (persistence/disappearance) of a certain social pattern, we are not able to provide all the relevant evidence that such an historical inquiry would ideally require;
3. offering a theoretical explanation of a social pattern, we want to cast light on its self-sustaining nature.

The point of this section was to show that Pettit’s idea of virtual self-interest is a helpful gloss on the practical, historical and theoretical variants of invisible-hand explanations. In the last section, we shall question the plausibility that the self-centred assumption is supposed to have recovered when it is construed as virtually operating.

**Virtual self-interest under scrutiny**

A key ingredient of Pettit’s resilient explanations is the notion that self-centred motivations have a virtual presence in deliberation. The precise sense in which self-centred motives are virtually at play can, however, be further specified. One way of doing this is to ask how self-centred motivations, virtually understood, relate to selfless ones. The following three claims, I submit, help specify such a relationship:

1. Selfless motivations are *the default* state of affairs and self-centred motivations deviations thereof.
2. Both self-centred and selfless motivations sustain the very *same* behaviour.
3. Virtual self-centred motivations and selfless motivations are *mutually exclusive*.

Let us elaborate on each of these three claims. One way of making sense of virtual self-centred motives is to say that they occupy a secondary position within the hierarchy of motivations. Consistent with Pettit’s view is the idea that selfless motives are prior motives in at least two senses. The first is methodological: self-centred motives should preferably be invoked as a second-best explanatory strategy (lack of evidence as to which motives really are at play was previously suggested as one rationale for adopting such methodology). Pettit seems to consider self-centred motives as secondary also in a more substantial sense. He says that self-centred motives start driving
choices only when selfless motives stop doing so. Selfless motives are the defeasible state of affairs, or, as Pettit says, the ‘normal mode of exchange’ (Pettit, 2000: 39), in reference to which the activation of self-centred motives represents occasional departures. Pettit metaphorically refers to both motivations in terms of the contrast between ‘automatic pilot’ and returning to manual. As he says, ‘self-regard will normally have no actual presence indicating what people do… The agent will run under cultural pilot, provided that that pilot does not carry them into terrain that is too dangerous from a self-interested point of view. Let such terrain come into view, and in most cases the agent will quickly return to manual’ (2000: 42). In sum, selfless motives form the baseline and self-centred ones are breaches of the latter. Consistent with this hierarchy of motives, resilient explanations confer on selfless motivations a privileged motivational role.

Pettit seems to be committed to the view that what requires an explanation is not the states where agents are moved by cultural, moral or civic ideals, but the circumstance in which they are no longer influenced by these ideals. We may object, however, that selfless considerations are given too important a role in daily life when they are construed, as Pettit does, as part of the status quo. It just does not sound accurate to say that we are routinely moved by cultural, moral or political duties. Surely, we also feel the latter as obligations we have no particular desire to follow.

Self-centred motivations are virtual in another sense: they would be the cause of agents’ choice if the latter were not motivated by selfless considerations. They are, in other terms, the pre-empted causes of behaviour. This is in turn compatible with never having any impact. Self-interest may endlessly remain the virtual, pre-empted cause of behaviour. Pettit seems to recognize the possibility of ascribing resilience to social patterns even in the absence of the actualization or manifestation of any such capacity when he says: ‘The picture leaves open the possibility that in many cases some individuals will not heed the alarms and will stick to what the culturally framed situation requires, by criteria of common values, through the thick and the thin of self-sacrifice’ (2001: 89). Consistent with this view is the possibility of seeing resilience as a mere dispositional property. Compare, in this regard, the property of resilience with that of fragility. No need to actually throw the glass against a brick wall to ascribe to it the dispositional feature of fragility. Merely being able to predict that it will break suffices. Analogously, it is thus possible to say of a social pattern that it is stable only by mentally submitting it to perturbing factors and imagining how it would react.6

Furthermore, Pettit’s theory applies only to cases where self-interested desires and selfless desires happen to sustain the very same behaviour. Only if deliberating in a self-interested manner produces no change of behaviour will it explain the resilience of the social pattern that is obtained in virtue of its virtual presence. As Pettit says, ‘self-regarding deliberation … leads … back towards the original pattern’ (Pettit, 2000: 45). ‘Self-interest’, Pettit also claims, ‘will kick in and stabilize the pattern’ (2000: 46). Deliberating self-interestedly thus needs to lead agents to remain faithful to their past choices and behaviour. The behaviour that agents display when they are moved by selfless motives will have to be the same as that which self-interested considerations require them to adopt. In fact, the metaphor of the little posts is consistent with this prediction:
Imagine a little set-up in which a ball rolls along a straight line – this, say, under Newton’s laws of motion – but where there are little posts on either side that are designed to protect it from the influence of various possible but nonactualized forces that might cause it to change course; they are able to damp incoming forces and if such forces still have an effect – or if the ball is subject to random drift – they are capable of restoring the ball to its original path. The posts on either side are virtual or standby causes of the balls’ rolling on the straight line, not factors that have an actual effect. (Pettit, 2000: 46)

The posts are not responsible for the rolling of the ball; nor do they explain the continuation of its straight course. They do, however, ensure that the ball ‘sticks to [a] more or less that straight line under the various possible contingencies where disturbance or drift appears’ (Pettit, 2000: 45). The posts on each side of the straight line are designed in such a way that they restore the ball to its original path.

One may reply, however, that deliberating in the light of self-interested concerns cannot be assumed to always have a restorative effect. Taking a look at the situation from the perspective of personal advantage may not systematically lead agents to stick to their habits. In fact, Pettit seems at times to recognize the possibility that deliberating in self-interested terms might end up in a change of behaviour. Sometimes, he admits, ‘the alarm bells will ring, causing the agents to rethink and probably reshape the project on hand’ (2000: 42; emphasis added). Self-interest may thus have a disruptive role in causing people’s behaviour to deviate from its usual path, giving rise to an entirely new form of behaviour. The posts in the metaphor need to be thought of as being capable of turning the ball away from, as well as redirecting it to, its Newtonian path. The bumpers of a pinball game, i.e. the round knobs that, when hit, push the ball away in all kinds of erratic directions, may here serve as a more suitable metaphor.

However, Pettit cannot consistently allow for the case where agents end up revising their behaviour as a result of consulting their self-interest. Allowing self-interest to have such a rectifying effect is not consistent with its recognized stabilizing power. A change of behaviour would put an end to the social pattern that had been obtained by virtue of agents’ past choices and behaviour. If the model of virtual self-interest is to explain the resilience of social patterns, cases where it leads to ‘reshaping the project at hand’ are ruled out. The model of virtual self-interest will explain the resilience of social patterns only if agents, when they deliberate self-interestedly, leave the project as it is. In other words, a reference to a virtual form of self-regard explains the resilience of certain social patterns under the following condition: the alarm bell rings, people consider whether or not what they are doing threatens their self-interest, only to realize that the way they have been behaving does not, after all, conflict with their personal interest (it was a false alarm, so to speak). In sum, we may say that self-centred motives will be virtually at play, in the specific sense Pettit has in mind, in the circumstance where they result in no change of behaviour.

The third sense in which selfless considerations are virtual is that they are simply ignored as long as the satisfaction of self-interest does not move below some threshold. Once such a threshold is passed, selfless motives cease entirely to drive agents’ choice, the latter being in the exclusive grip of self-centredness. Selfless and self-centred considerations are, in this particular construal, assumed to be mutually exclusive. Both motivations have an independent, asynchronous bearing on choices, the simultaneous presence of both motives not being a possibility that Pettit considers.
This view remarkably excludes the trivial possibility that agents routinely consult and compare both their selfless and their selfish preferences when they deliberate about what they should do. It rules out the case where we weigh moral and cultural considerations along with and against self-interested considerations. To be sure, we may easily grant that self-interested considerations can come to outweigh others, those circumstances being when self-interest is not fully maximized. But self-interested considerations need not chase away selfless ones when they enter the picture. Nor need the latter have an exclusive presence in agents’ minds when they actually govern their choice and decision. The two kinds of considerations could indeed coexist, albeit presumably not peacefully, within one’s mind. This latter view is arguably a more plausible way of considering the way altruistic and selfish considerations respectively bear on our choice than Pettit’s ‘all or nothing’ view. Here we may want to say that, while proponents of self-centred based explanations go wrong in seeing selfishness everywhere, Pettit may go too far in denying it any current, psychological reality.

The upshot of the discussion is that there are various constraints that self-centred motivations must meet to play the specific virtual role that will explain resilience rather than emergence or continuation. We have argued in particular that they need, first, to constitute a departure from the normal situation in which selfless motivations allegedly bear on our choices. Second, they will have to sustain the same behaviour. And, third, they will have to be absent when the former are present. As it should appear clearly now, the conditions under which virtual self-interest explains the resilience of social patterns turn out to be both cumulatively and separately quite restrictive.

Conclusion

The aim of this article was, first, to critically assess invisible-hand explanations in light of Pettit’s model of virtual self-regard. The idea was to examine to what extent their various cognitive worth could be elucidated by means of such a model. Those who offer explanations of this type may thus simultaneously pursue various goals. One is to point to the desirable way of obtaining certain desirable social outcomes that dispenses with the difficulty of coordination that agreement raises, on the one hand, and the risk of having to count on unreliable legislators, on the other hand. Another is to give an account of the way an institution may have arisen in the absence of the possibility of knowing the exact details of its actual emergence. Still another purpose is to offer a theoretical account of the functioning of an institution (like the money system, property rights or the institution of the state) that abstracts from the particular settings of its various instantiation. We have shown how each of these aims can be suitably pursued by assuming, à la Pettit, that agents are likely to approach their situations in self-centred terms.

While a better understanding of this particular type of explanations can be reached by means of Pettit’s model, we also have given reason for thinking that the explanatory value of their highly abstract, theoretical variant may not be entirely resolved in this manner. In our critical discussion of Pettit’s idea of resilience explanation, we show that the idea of virtual self-regard is still at variance with the way we commonly take the mind to ordinarily work in most social settings. This may show that, pace Pettit, making self-interest a concern of the virtual kind may not fully protect economic
explanations, such as invisible-hand ones, from what he refers to as ‘the problem of the empty black box’.

**Acknowledgements**

I would like to thank Uskali Mäki and two anonymous referees for their helpful comments and criticisms on previous versions of this article.

**Declaration of conflicting interest**

The author declares that there is no conflict of interest.

**Funding**

This article was written with the support of the Marie Heim-Vögtlin program of the Swiss National Foundation (project research PMPDP1_145534)

**Notes**

1. Bunge blames invisible-hand explanations for their ‘oversight of the social context of agency … bound to lead to utterly false models of social facts, as well as to immoral social policies’ (Bunge, 1997: 440). Elster stresses the confusion that explanations that are based on implausible behavioural assumptions are likely to elicit when their speculative nature is not fully recognized. ‘Storytelling can be harmful’, he says, ‘if it is mistaken for the real thing’ (Elster, 1989: 7). Invisible-hand explanations are, according to Rosenberg, ‘empty and untestable post hoc reasoning’ (Rosenberg, 1995: 161).

2. The explanatory worth of invisible-hand explanations, Aydinonat argues, is their capacity to ‘expand our mental horizon… indicating new paths for further empirical and theoretical research’ (Aydinonat, 2008: 107). Nozick says about the invisible-hand explanation he provides of the minimal state that ‘even if no actual state ever arose that way’, it will ‘serve our explanatory purposes almost perfectly’ (Nozick, 1974: 7). Invisible-hand explanations, according to Ullmann-Margalit, ‘yield substantial understanding of, and provide considerable insight into, the nature of the social patterns and institutions explained, and … are of remarkable explicatory worth’ (Ullmann-Margalit, 1978: 277). ‘One may claim that an invisible-hand interpretation is of interest’, Vanberg suggests, ‘even in case [the institutions to which it is applied have been deliberately established] because it can provide information about the general nature of the institution and the forces that support it’ (1994: 151).

3. Adam Smith is the first to metaphorically refer to an invisible hand when he discusses the advantage of not having to rely on a central authority for wealth to be redistributed (Smith, 1982 [1759]). Earlier occurrences of the twin notion of desirable unintended consequences are Mandeville’s *Fable of the Bees* (1988 [1732]; see Hayek, 1966), even if the details of the various mechanisms by which ‘private vices’ transform into ‘public benefits’ confers a too crucial role to ‘skilful politicians’ for being genuinely illustrative of the invisible hand (Viner, 1958). Hume’s evolutionary theory of the emergence of property rights (Hume, 2000 [1740]; Williams, 2002, chap. 2), Menger’s explanation of the emergence of money (Menger, 1892) and Nozick’s explanations of the spontaneous institution of a minimal state (1974) are often cited as paradigm examples. While Nozick first coined the term ‘invisible-hand explanations’, Ullmann-Margalit’s reference article (1978) remains perhaps the most comprehensive conceptual clarification of the notion (see also Aydinonat, 2008; Brennan & Pettit, 1993; Tieffenbach, 2013).
4. The idea that agents are blind to what they do is often considered as a core ingredient of all invisible-hand explanations (see for example Nozick, 1974; Pettit, 1998; Ullmann-Margalit, 1978). This constraint can be understood in many ways. Unintended consequences can, first, be taken to be not only no one’s goal but also effects that are not even foreseen. Alternatively, it may be sufficient that the outcome to be explained be merely no one’s goal to be labelled as unintended, leaving the possibility of these consequences to be anticipated. Still alternatively, the notion of unintended consequences may apply to outcomes that could be both individually anticipated and brought about on purpose, but not as a result of a joint action. This last interpretation, as I have argued elsewhere, could well be the most coherent one (cf. Tieffenbach, 2013).

5. Another way of applying the invisible hand to a normative goal is to have it perform a justificatory role (see Schmidtz, 1990). This time, the idea is to demonstrate the merit of the outcome to which the invisible-hand process leads, rather than that of the path leading to it. Showing that some outcome could be the ultimate stage of an invisible-hand process indicates its goodness or desirability. By many accounts, Nozick’s explanation of the minimal state is illustrative of this justificatory use. Akin to the social contract hypothesis, the invisible hand is then called for with a view to offering a justification of what it brings about. Agents that are led by the invisible hand do not, by hypothesis, violate anyone’s right (such a virtuous disposition seems to be an ad hoc requirement), and thus contribute to the establishment of an outcome that will inherit the legitimacy of the path by which it has arisen. The reason we make the path a hypothetical one, rather than an historical one, is that we want its legitimacy not to be a matter of what fallible agents actually deliberate and act upon, but a matter of what their ideal counterparts would do.

6. On an alternative conception, social dispositions are more akin to psychological dispositions than to physical dispositions. Just as we would not say of a person who has never had the opportunity to demonstrate her bravery that she is brave, we would not say of the money system that it is a robust system merely on the basis of its imagined reaction to a shifting of behavioural motivations on the part of those who use it. Dispositional institutional features may well behave like their psychological counterparts, but they need to manifest themselves once in a while in order to be so ascribed. But if this view is correct, it also means that they can be attributed only ex post.

References
Tieffenbach


**Author biography**

After a PhD in philosophy on invisible-hand explanations (University of Geneva), Emma Tieffenbach pursued her research in social ontology as a post-doctoral researcher at the Swiss Centre for Affective Sciences in Geneva. Her current project, under the benefit of a Marie Heim-Vögtlin subsidy from the Swiss National Foundation, deals with the subjectivity of economic values and the nature of exchanges.