REFLEXIVITY AND EQUILIBRIA

FRANCESCO GUALA

Working Paper n. 2013-16
SETTEMBRE 2013
Reflexivity and Equilibria

Francesco Guala*

Abstract
The failure of models based on rational expectations to explain the “boom and bust” of financial markets does not support Soros’ critique of mainstream economics or his call for a theoretical revolution. Contrary to what Soros says, standard rational choice theory has the conceptual resources to analyse reflexivity. The dynamic of feedback loops for example can be described by simple models based on multiple equilibria and informational cascades. The problem is that agents and theorists sometimes lack the information required to identify equilibria and tipping points.

JEL Classification: B41, C70

Keywords: reflexivity, equilibria, informational cascades

* Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, via Conservatorio 7, 20122 Milano, Italy. Email: francesco.guala@unimi.it.
1 Introduction

“Reflexivity” is a popular term in scholarly conversations. The fact that it is used in different fields has certainly contributed to its diffusion, but at the same time has generated confusion. Its core meaning comes from mathematics: a reflexive relation (“is equal to”, for example) relates every element of a set to itself. In social theory reflexivity has been mainly used in the context of causal, rather than logical or mathematical relations: in a reflexive causal relation, each event or variable is both a cause and an effect of another event or variable (X causes Y and Y causes X).

Most accounts of reflexivity emphasise the role of beliefs. In George Soros’ words, “the participants’ views influence but do not determine the course of events, and the course of events influences but does not determine the participants’ views” (2013: 9). An attraction of the term “reflexive” is its relation with the phrasal verb “to reflect on”, meaning “to think carefully” on a given topic. Social phenomena are reflexive because people are reflexive in this sense.

Soros argues that economists’ failure to predict and prevent economic crises is a consequence of economic theory’s neglect of reflexivity. He also claims that a theory that takes reflexivity seriously will depart radically from mainstream models, causing a paradigm shift from current economic thinking. Although I agree with Soros’ on the importance of reflexivity, I think that his critique of mainstream economics is mistaken. Reflexivity plays a central role in economic thinking, and rigorous models based on this concept are unlikely to depart substantially from standard models based on Nash equilibria and equilibration processes.

The paper is structured as follows: in section 2 I explain why the concept of reflexivity, as understood by Soros, is strictly related to the concept of equilibrium used in rational choice theory. Section 3 provides a simple example of how reflexivity, as well as “negative” and “positive” feedback loops, can be represented using simple models of equilibrium and informational cascades. Section 4 extends the discussion arguing that reflexivity and equilibria are key concepts to represent not only “boom and bust” phenomena, but also the stability and persistence of social institutions. Section 5 summarizes and concludes the paper with some reflections on the future of economics.
2 Equilibrium

George Soros’ views on reflexivity originate from his life-long study of financial markets. Reflexivity is a likely source of the disruptive “boom and bust” cycles that unsettle financial markets from time to time, and economists and politicians do not seem to have found a way to immunize the economy from these cycles so far. So Soros’ emphasis on reflexivity and his call for increased research effort are certainly welcome. Unfortunately Soros’ misdiagnoses the causes of economists’ failure, and as a consequence overstates their implications.

One source of misunderstanding is Soros’ conflation of rational expectations theory and the Efficient Markets Hypothesis (EMH) with economic theory. The EMH is not representative of economics as a whole: it is a particular, and highly contentious, application of the theory of rational choice that lies at the core of most neoclassical economic models. As a mere point of logic, the failure of EMH can hardly be imputed to economics in general.

Rational expectations models assume that market participants have correct (rational) beliefs about the main economic variables, and as a consequence that market prices do not diverge significantly from equilibrium. But such claims rely on the crucial underlying assumption that there is a unique equilibrium and that the equilibrium is efficient. Although such assumptions are ubiquitous in finance, other branches of economics routinely use models with multiple equilibria, some of which are inefficient. These models are well suited to capture the dynamics of reflexivity that Soros is concerned about, because equilibrium is itself a reflexive notion.

In the technical jargon of economic theory, a Nash equilibrium is a profile of actions such that the action of every individual is optimal given the actions of the other individuals: each action is a best reply to the actions of others. This does not mean, as every economist knows, that the overall outcome (the profile of actions) is efficient. It just means that no one can achieve a better outcome by changing her strategy unilaterally. An important property of this concept is that the beliefs of all individuals are true in equilibrium. I choose X because I believe (correctly) that you choose Y, and you choose Y because you believe that I choose X. Another way to put it is that the beliefs of all the agents form a gigantic self-fulfilling prophecy.
Soros describes this mutual consistency of actions and beliefs informally, using the concepts of “manipulative function” and “cognitive function” of human thinking. Roughly, these refer to the fact that human agents simultaneously try to represent the world but, through their actions (which are influenced by their representations) also determine the state of the world they are in. According to Soros this mutual influence raises a theoretical puzzle:

when both the cognitive and manipulative functions operate at the same time they may interfere with each other. How? By depriving each function of the independent variable that would be needed to determine the value of the dependent variable. The independent variable of one function is the dependent variable of the other, thus neither function has a genuinely independent variable. (2013: 6)

Before Soros other social theorists have been baffled by this apparent circularity. Oskar Morgenstern gave one of the earliest formulations of the problem in his 1928 dissertation on Wirtschaftsprognose (economic forecast). One of his famous examples concerns two individuals – Holmes and Moriarty – playing a game of “hide and seek” (Morgenstern 1928: 98). If Holmes decides to take the train to Dover, Moriarty will try to catch him at the terminal station; Holmes should anticipate this, and stop at an intermediate station; but a rational Moriarty should anticipate Holmes’ anticipation, prompting Holmes to consider traveling all the way to Dover again, … and so forth for an infinite number of iterations. Morgenstern found this circularity – the fact that any move will provoke a symmetric defeating counter-move – extremely disturbing, and used it to draw pessimistic conclusions for the predictive ambitions of economic theory.

But economists have gone a long way since then, partly (ironically) thanks to Morgenstern himself. Any student of game theory now knows that seemingly unstable interactions such as these can and often do converge to an equilibrium (von Neumann and Morgenstern 1944). Take the game of penalty kicks, a situation that mirrors the Holmes-Moriarty game and that is routinely played by strikers and goalkeepers on European football pitches. Contrary to Morgenstern’s intuition, such games have a perfectly rational and stable solution (an equilibrium) according to which the strikers shoot sixty per cent of the time on their strong
side and the keepers dive sixty per cent of the time on the same side. This is a *mixed-strategy* equilibrium, a pattern that is statistically predictable, is compatible with individual rationality, and leaves a lot of uncertainty and fun for the fans who watch the game on TV.

The concept of equilibrium, as used in strategic models, has allowed rational choice theorists to tackle these reflexive dependencies in an analytical way. At times Soros seems to accept that equilibrium must play a role in a proper theory of reflexivity. But the latter, he argues, will have to go much beyond equilibrium, in order to provide adequate understanding: “equilibrium, which is the *central* case in mainstream economic theory, turns out to be an extreme case of negative feedback, a *limiting* case in my conceptual framework” (2013: 29).

So the problem is that economics does not go far enough. Its narrow focus on a limiting case prevents it from giving an adequate account of *out-of-equilibrium* behaviour. This is what Soros has in mind when he says that “reflexive feedback loops have *not* been rigorously analyzed” (2013: 9). But what is it exactly that economists are missing, and how would a better theory look like? Soros is not shy and tries to give an informal account of reflexivity. His stab is based on the twin notions of negative and positive feedback: “negative feedback brings the participants’ views and the actual situation closer together; positive feedback drives them further apart” (2013: 29).

Now, is economics really impervious to such notions? In rational choice terms, a negative feedback loop is a process of belief revision that converges toward a Nash equilibrium, while a positive feedback loop moves away from it. In the next section I illustrate a simple model of these processes. The model is abstract, but to make it concrete we can dress it with an interpretation taken from Robert K. Merton’s classic discussion of the self-fulfilling prophecy (Merton 1948).

### 3 Equilibration

Merton’s story begins with a bank that is financially sound. The bank has plenty of liquidity for its day-to-day business and no prospect of insolvency in the foreseeable future. One day, however, an unusual number of depositors appear at its desks to withdraw their savings. To

---

1 Cf. Palacio-Huerta (2003). The “strong side” of a striker is left for right-footed players, and right for left-footed players.
make sense of this unexpected event, some other clients conjecture that the bank is in trouble, and decide to withdraw their money too. This causes more cueing in front of the desks and more withdrawals. Very soon the rumour becomes true – the bank is insolvent and goes bankrupt.

Merton highlights the common pattern in this and other cases of reflexivity: a false belief brings about events that turn it into a true belief. The bank was not unsound before the rumour spread. But because people came to believe it, the bank became unsound. As Soros would say, this is the “principle of fallibility” in action.

To Merton, self-fulfilling prophecies are obvious aberrations and raise two sets of concerns. First, there is a policy problem: can we prevent them from occurring? It seems that we would all be better off if our financial system was invulnerable to the diffusion of unfounded rumours. But there is also a scientific problem: self-fulfilling prophecies are puzzling – they are “peculiar to human affairs”, “not found in the world of nature”, according to Merton – so we ought to find an explanation.

Although his essay remains a classic reference, Merton did not propose a satisfactory analysis of the mechanics of self-fulfilling prophecies. In the two decades that followed, however, economists devised simple tools to model such situations (Figure 1). To simplify, let us assume that each depositor can only choose one of two actions: to keep her money stored in the bank, or to withdraw it. On the horizontal axis we represent the proportion of depositors who are expected to withdraw, on a scale that goes from zero to one. For simplicity, we interpret this parameter as an aggregate function of the individual beliefs of all depositors, something like the “average” expectation in the population. On the vertical axis, we represent the proportion of depositors who actually withdraw (again on a zero to one scale). The diagonal line includes all the points where the expected number of withdrawals is equal to the actual number of withdrawals – that is, the set of all possible correct beliefs.
The S-curve in the diagram is called the propagation function, because it represents how many people decide to withdraw, given their beliefs about the proportion of withdrawals. If we suppose that customers observe in real time each other’s behaviour, we obtain a dynamic model known as a “Schelling diagram”, from the game theorist and Nobel Laureate Thomas Schelling (1978).

The propagation function in Figure 1 crosses the diagonal in three different points: $e_1$, $e_2$, and $t$. These are self-fulfilling prophecies: people’s beliefs about withdrawals (and, a fortiori, the soundness of the bank) are correct. A dynamic model however allows more than simply identify self-fulfilling prophecies and beliefs. It also permits an analysis of the forces that lead individuals to adjust their behaviour as they update their beliefs. This analysis shows that over time some outcomes are more likely to occur. As Merton pointed out, certain self-fulfilling prophecies are very predictable – indeed, they appear inexorable. Looking at the model we can understand why.

Let us start from the bottom-left tip of the S-curve. In this corner of the box we have those depositors who are going to withdraw their money no-matter-what. We can imagine these clients as people who need the money for some expense or investment, which was planned independently of any turmoil in the financial sector. Even if no one else withdraws, they will do it. All the other depositors instead condition their choices on their expectations regarding the proportion of withdrawals. Now, when the propagation function lies above the diagonal,
we have a situation where more people are withdrawing (vertical axis) than it was expected (horizontal axis). There is a contradiction between beliefs and behaviour that must be resolved. Once the lines begin to form in front of the bank, depositors process the new information and revise their beliefs accordingly. This corresponds to a shift towards the right in the diagram: a few more clients decide to withdraw their money from the bank. If we are still above the diagonal, the same sequence of observation, belief-revision, and withdrawal continues until the propagation function meets the diagonal line.

Notice that when the reaction function lies below the diagonal, we have an adjustment in the opposite direction. Fewer people withdraw their money than expected. Again, customers process this new information and revise their beliefs accordingly, causing a shift towards the left-hand side of the box. These two dynamics jointly imply that the adjustment process tends towards \( e_1 \) or \( e_2 \) (see the arrows in the figure). Soros describes this equilibration process informally in these terms:

> A negative feedback process is self-correcting. It can go on forever and if there are no significant changes in external reality, it may eventually lead to an equilibrium in which the participants’ views come to correspond to the actual state of affairs (2013: 28-9).

He also points out that the system does not always return to the original equilibrium, though:

> Positive feedback loops are more interesting because they can cause big moves both in market prices and in the underlying fundamentals. A positive feedback process that runs its full course is initially self-reinforcing in one direction, but eventually it is liable to reach a climax or reversal point, after which it becomes self-reinforcing in the opposite direction. But positive feedback processes do not necessarily run their full course; they may be aborted at any time by negative feedback. (2013: 30).

This indeterminacy in the outcome of feedback processes is explained in the model by point \( t \). By definition, \( e_1, e_2, \) and \( t \) are all self-fulfilling prophecies. Beliefs and behaviour are always consistent on the diagonal. But \( e_1 \) and \( e_2 \) are robust, in the sense that small deviations from these points will trigger processes that bring the system back to its previous state. Point \( t \) in contrast is very fragile: it is a “tipping point”, meaning that even the slightest deviation
will start a dynamics that will make the system converge towards $e_1$ or $e_2$. Social scientists call these dynamics *informational cascades*, and use them to explain abrupt institutional changes such as political revolutions (Kuran 1995). When a tipping point is crossed, there is no coming back – the gulf will increase and the fate of the bank is sealed.

How can it possibly happen? What triggers a run on a bank? Some event must push beliefs beyond the tipping point. If the system normally tends to gravitate around $e_1$, such an abrupt change can be brought about only by an external shock. This may be a new piece of information, such as for example a report from a rating agency saying that the bank has falsified its accounts. Other abrupt events that may have the same effect include a declaration of war, or a sudden political crisis. The common feature of such events is that everyone has to form very quickly new beliefs about the behaviour of others, in unusual circumstances that they have never faced before. There is a lot of uncertainty, and people in such moments often go for the option that seems the safest – to withdraw their savings for example. In these situations of panic, the whole system may crumble very quickly, and it is difficult to intervene before it is too late.

When beliefs are fluid, economic theory can have a stabilizing function. If the beliefs of a sub-class of individuals (economists or market analysts) enjoy a particular epistemic authority, they will tend to influence the views of other participants. This process of imitation is analogous to the dynamics described above, and may direct behaviour towards or away from a specific equilibrium. This is the sense in which social theorizing can influence reality, a feature that many philosophers – including Soros – consider truly distinctive compared to natural scientific theorizing.\(^2\)

4 Social reality

Soros does not consider the possibility of multiple equilibria. His speaking of an “objective reality” that “reasserts itself”, or of “far-from-equilibrium situations” that persist over time, suggests that he is in the grip of a key mainstream assumption: markets have a unique equilibrium. Once this assumption has been taken on board, feedback processes do appear puzzling and radically at odds with economic theorizing. But if multiple equilibria and

---

\(^2\) McKenzie’s (2006) narrative of the rise and fall of the Black-Scholes model of option pricing is based on this idea and provides a neat example of reflexivity in action.
tipping points are admitted, much of the mystery disappears. The problem becomes *practical* – how to model complex feedback processes adequately – rather than conceptual.

The practical problem is daunting, and explains social scientists’ failures to forecast political and economic crises. The shape of the propagation function is usually unknown to agents and theorists alike. This implies that they are typically unsure about the existence and location of equilibria and tipping points. In times of uncertainty, forecasting becomes almost an impossible task. And reform, for the same reasons, is extremely difficult when people doubt that a superior equilibrium can be attained. Many aspects of the Euro crisis can be explained this way.

Is such a predicament inevitable for social science? Soros claims that equilibrium is “an extreme condition”, and stability is “exceptional” (2013: 28). “The human uncertainty principle implies that a perfect alignment would be the exception rather than the rule” (2013: 17). I think this is a misunderstanding, prompted perhaps by the uniqueness of equilibrium assumption. The truth is the opposite: equilibrium or near-equilibrium is the norm in social life. What may look like the persistence of far-from-equilibrium behaviour is often convergence toward another equilibrium that we had failed to consider before.

Equilibrium is the norm because radical uncertainty would make social life impossible. Human societies are enormously complex systems of coordination games, and coordination requires stability of behaviours and beliefs. This is not to deny the existence of coordination failures, or of systemic crises that occasionally affect large portions of society at once. During these periods of crisis we tend to emphasise the fragility and unpredictability of the structure, forgetting the enormous resilience and consistency of expectations that makes social life possible. Even simple actions like purchasing bread at the local baker’s would be impossible if far-from-equilibrium behaviour was the norm. Crossing the street would be a heroic feat without the equilibria of the rules of traffic; bartering with the baker would take an enormous amount of time without the equilibrium of a common currency.

Emphasising the fragility rather than the robustness of social systems is like seeing the glass half empty or half full. This is why reflexivity elicits opposite reactions. Understandably, it has often been perceived a source of problems to be solved: if you are a scientist interested in explanation and prediction, reflexivity makes your life harder. But some scholars rejoice and
celebrate feedback loops – like Douglas Hofstadter in his best-selling book on *Gödel, Escher, Bach* (subtitle: *An Eternal Golden Braid*). Loops may have creative, generative power. They may bring new kinds and phenomena into being. They may even bring about new ways of living – new human identities, as Ian Hacking has pointed out.³ In general, feedback loops need not be disruptive.

Of course stability is not necessarily good in itself. There are different equilibria in our simple model of a bank run, and one of them (\(e_1\)) is intuitively better than others: it is the equilibrium where the bank exists and it is financially sound. The situation in \(e_2\) is quite different: although everyone is “happy”, in the sense that no one would like to change her behaviour unilaterally, each depositor regrets the demise of the bank. So it is possible that each individual is trapped in \(e_2\) even though she recognizes that she would be better off in \(e_1\).

The model of the run on a bank provides an important insight that should inform any social theory. The insight is that both states \(e_1\) and \(e_2\), where the bank is respectively sound and unsound, are self-fulfilling prophecies. A shift from \(e_1\) to \(e_2\) appears perverse because the new beliefs do not seem to reflect a change in the “fundamentals” of the bank. But this is misleading: deposits are one of the fundamentals. So it is not irrational to withdraw, if the other customers are withdrawing. Since in \(e_2\) all beliefs are correct and all preferences are satisfied (given these beliefs), the difference between \(e_1\) and \(e_2\) is not as deep than it may initially appear.

Another way to put it is that there is nothing weird or “wrong” about reflexivity itself. Self-fulfilling prophecies may lead to equilibria that we like, as well as equilibria that we do not like (it depends on our preferences). But without reflexivity or self-fulfilling prophecies a lot of important institutions that we love and respect would not exist. There would be no banks, for sure, but also no political leaders, no religious ceremonies, no football teams, and so on and so forth. It took a long time for this simple truth to be appreciated by social scientists and philosophers.⁴ The game-theoretic perspective on social institutions, which was developed independently and for other purposes, facilitates the construction of simple models of reflexivity. What is a sound bank? In equilibrium, I do not withdraw my money because I

---

³ See e.g. Hacking (1985, 1999), Mallon (2007).
⁴ As far as I know, it was first noticed by the philosopher Daya Krishna, in a 1971 commentary on Merton’s essay. For an influential early analysis of reflexivity along these lines, see also Barnes (1983).
believe that you do not withdraw, because I believe that you believe that I do not withdraw, and I believe that you believe that I believe that you do not withdraw, and so on. Social reality is constituted by beliefs about beliefs.

5 Conclusion

Soros’ reflections are driven by two different claims. One is of them correct: reflexivity is a key concept to understand the dynamics of market behaviour and of social phenomena more generally. The second thesis is that economic theory does not have the conceptual resources to deal with reflexivity. As I have shown in this paper, this claim is misleading. In part, it is based on the conflation of economic theory with a specific application, the EMH. The EMH is a particularly narrow rational choice model, and its failure does not have any implications for the prospect of economics to explicate reflexivity. As we have seen, the phenomena that Soros classifies as “reflexive” can be easily represented using rational choice models such as Schelling’s diagrams. The strategic interactions depicted in these models are admittedly simple, and economists face enormous difficulties when they try to explain the functioning of more complex systems such as real financial markets. But the very existence of these models demonstrates that the challenge is practical, rather than conceptual, in character. There is no deep flaw in economic theory that prevents it from capturing the essential features of “negative” and “positive feedback loops”. Soros claims that “the new paradigm [based on reflexivity] is bound to be very different from the one that failed” (2013: 41). I think that the opposite is true: any theory that will succeed to model such phenomena will have to employ concepts that are very similar to equilibria and equilibration processes.

Once this has been clarified, other mistakes are easily avoided. Soros follows a long series of social theorists in identifying “physics’ envy” as a fundamental obstacle for the progress of social science. But physicists and biologists make extensive use of equilibrium models, and equilibrium is a reflexive notion. So physics’ envy cannot be the problem. That neoclassical economic theory borrowed some analytical tools from Newtonian physics is a historiographic platitude. Overall this may have been a bad or a good idea, but it has certainly not prevented economists from understanding reflexivity (if anything, the opposite is true). Newtonian physics did not dictate economists the theory of rational expectations or the EMH. Financial

---

5 See e.g. Ingrao and Israel (1987), Mirowski (1989).
economists “ended up” with such hypotheses for a number of contingent reasons, but their ignorance of reflexivity was not one of them. In fact the central modelling tools of rational choice theory, such as equilibria and belief revision, capture the essential features of reflexivity remarkably well. If reflexivity is a fundamental feature of social phenomena, as Soros contends, then the economic theory of choice is well equipped to explain social reality.

Of course this does not demonstrate that we should stick to rational choice theory. Perhaps we should really change strategy and abandon traditional models based on equilibria and equilibration processes. After all, we might never be able to understand, control, and predict the behaviour of financial markets using such models. Perhaps a Copernican revolution is needed in this domain. But – paradoxically for Soros – any theory that will abandon equilibrium concepts will also abandon reflexivity. Because reflexivity and equilibrium models are so strictly related, they will stand or fall together in any future social theory.

Acknowledgments: This paper was written while I was supported by a MIUR grant “Rientro dei Cervelli”.

References


