Public Reason 10 (1): 33-48

Rawlsian Stability and the Hazards of Envy

Alexandros Manolatos
National and Kapodistrian University of Athens

Abstract: This paper explores the role of envy in the third part of *A Theory of Justice* and challenges a wide-spread game-theoretic view of stability. The proponents of this view see Rawls's account of stability as an attempt to solve a collective action problem. I claim that Rawls treats the development of envious feelings as a distinct source of instability which is not part of a collective action problem and has to be addressed separately. My thesis entails that we shouldn't read the congruence between the right and the good as the culmination of Rawls's overall argument for stability. This reading is supported by the revised account of stability in *Justice as Fairness: A Restatement* and leads to a better understanding of Rawls's political turn.

Key words: Rawls, stability, envy, game-theory, Weithman, congruence.

The problem of stability is according to Rawls fundamental to political philosophy. A political conception of justice must be utopian and at the same time realistic. It must contain principles and ideals that can be seen as possible, given the circumstances of justice and the laws and tendencies of the social world.

The central role of stability in Rawls's political theory is reflected in the structure of *A Theory of Justice.*[1] The third part of the book is devoted to an examination of the questions of stability and congruence. As Rawls (1971, 55) writes in *TJ*, however attractive a conception of justice might be, it is defective if it fails to engender in human beings a desire to act upon it. So the main aim of the third part of Rawls's magnum opus is to show that the principles selected in the first part are going to generate their own support and are more stable than other principles. This is shown by the assessment of the relative strength of the sense of justice they cultivate and of the opposite temptations to act unjustly.

The significant role of stability in Rawls's political philosophy is reflected also by his turn to *Political Liberalism.*[2] In the introduction to *PL* he informs the readers that he decided to revise his theory because the account of stability he presented in *TJ* was not realistic enough. These revisions were necessary because "the problem of stability is fundamental to political philosophy and an inconsistency there is bound to require basic readjustments" (Rawls 2005, xvii).

Surprisingly, even after the publication of *PL*, Rawls's account of stability remained initially at the margins of the overall academic discussion on his work. Freeman (2007) has remarked accurately that from all the voluminous commentary on Rawls's work, very little of significance has been written on his argument for stability in the third part of *TJ.*[3]

[1] All references to A *Theory of Justice* are to the original edition, unless otherwise stated. The book is cited hereafter as *TJ*.

[2] All references to *Political Liberalism* are to the expanded edition. The book is cited hereafter as *PL*.

[3] Freeman made that remark almost ten years after the publication of *PL*. He was one of the first who stressed the importance of stability in Rawls's work.

One possible explanation is that most critics saw Rawls's political turn as a response to the critique of communitarianism.[4]

The first significant exclusive commentary on Rawls's argument for stability was Paul Weithman's book *Why Political Liberalism? On John Rawls's Political Turn*. Weithman offers a thorough and novel analysis of the third part of *TJ,* which reveals the role of ideals in Rawls's argument for stability. The publication of his book revamped the debate on Rawls's political turn and sparked an interest in the role of stability in his political philosophy. At the same time, Weithman's book established a game-theoretic view of Rawls's account of stability in *TJ*.

In this article my main goal is to oppose this game-theoretic view and to stress the importance of the special psychologies in Rawls's overall stability argument.[5] I will claim that the game-theoretic view distorts how Rawls conceives the problem of stability in *TJ,* since it focuses solely on the congruence argument and the challenges of mutual assurance and isolation. I will first outline the game-theoretic view of Rawlsian stability and then I will propose a different reading that gives equal weight to the problem of envy. This reading reveals the inherent connection of stability with the issue of distributive justice and social status. Envy is a serious challenge for the justification of justice as fairness because the inequalities sanctioned by the difference principle are theoretically unlimited and they could arouse hostile feelings that threaten social unity. A large part of Rawls's account of stability aims to show that this challenge can be met by the reciprocal and egalitarian character of the two principles. If we see the problem of stability only through the lens of game theory we lose this important aspect of Rawls's view.

In the final section of the article, I claim that if we abandon the game-theoretic view we can have a better understanding of Rawls's political turn.[6]

I. THE GAME THEORETIC VIEW OF RAWLSIAN STABILITY

In *TJ* the problem of stability is examined in the third part of the book, which comprises of three chapters. In the first chapter Rawls presents his theory of the good and the Aristotelian principle. In the second chapter he discusses how the citizens of a well-ordered society acquire a sense of justice and he compares the stability of the two principles of justice with other conceptions. The final chapter examines whether the sense

[4] Freeman notes that there is a widespread perception that the revisions Rawls has made to *TJ* leading up to *PL* have come largely in response to communitarian criticisms (2007, 175).

[5] In *TJ* Rawls presents envy and spite as special psychologies or special attitudes that limit rational behavior. The will for domination and submission and a peculiar aversion to risk and uncertainty are the other special psychologies he cites in *TJ*.

[6] In a series of papers between 1980 and 1987 Rawls started modifying his theory and revising some of his arguments. These revisions led to the publication of *PL* and the recasting of justice as fairness as a political and not as a comprehensive conception of justice. In the section "The Revised Account of Stability" I will discuss in more detail the main differences between *TJ* and *PL*.

of justice coheres with our good. The three chapters are interlinked, but the way the overall argument is structured is not so straightforward and has to be reconstructed.

One possible reconstruction is based on the assumption that Rawls conceives the problem of stability in game-theoretic terms.[7] The proponents of this game-theoretic view interpret the problem of stability in *TJ* as a collective action problem that arises from two sources of instability which threaten a well-ordered society.[8] The first source of instability arises from the fact that citizens of a well-ordered society will respect the two principles of justice only if they have the assurance that others will act in the same way. This is called by Rawls "the assurance problem" (1971, 270). The second source of instability arises from the fact that if they know that others comply with their duties, they will be tempted to ignore their duties and act as free-riders.

Those who support this view hold that Rawls's argument for stability is targeted exclusively at solving a generalized prisoner's dilemma and the related problem of assurance. They claim that the goal of the third part of *TJ* is to show that a well-ordered society would not suffer from these two sources of instability and that this is achieved by the congruence between the right and the good.[9] The congruence argument is presented in the section "The Good of the Sense of Justice". In this section Rawls argues that citizens of a well-ordered society would affirm their sense of justice because in that way they could satisfy four desires. The desire to express themselves as free and equal rational beings, take part in forms of social life that call forth their own and others talents, avoid the psychological cost of hypocrisy and have ties of friendship. The advocates of the game-theoretic view see this

[7] The influence of game theory in *TJ* is not evident only in the argument for stability. Rawls uses a general analogy between society and games which helps him emphasize the idea of fairness (Galisanka 2017). In *TJ* Rawls refers to this analogy when he discusses the rule of law (1971, 235), the idea of pure procedural justice (1971, 85), the idea of social union (1971, 525) and when he compares institutions with games (1971, 55). There is also an input of game theory in the argument for the selection of principles from the original position, where Rawls introduces the maximin rule (1971, 152). I would like to thank an anonymous referee for constructive comments regarding the overall use of game theory by Rawls.

[8] This view is expressed by McCLennen (1989), Weithman (2010), Thrasher and Vallier (2015), Quong (2014) and Freeman (2003, 2007). Freeman and Weithman have contributed decisively in illuminating Rawls's distinctive understanding of stability. Freeman (2007) reveals how the congruence argument is connected with the Kantian interpretation of justice as fairness. In his seminal book on Rawls's political turn, Weithman (2010) highlights the role of ideals of conduct, friendship and association in the justification of the congruence argument. I find these reflections convincing and I do not intend to challenge them. My objection centers on the fact that they both see the congruence argument as a culmination of the overall argument for stability.

[9] John Thrasher and Kevin Vallier (2015) contend that in *TJ* Rawls attempts to solve the assurance problem by showing that citizens have reason to endorse their sense of justice as part of their good, namely that the right and the good are congruent. This congruence ensures that "each person in the well-ordered society will be motivated by a sense of justice to comply with the public conception of justice and will know that everyone else is motivated in the same way" and this "common knowledge of compliance" preserves mutual assurance between citizens (Thrasher and Vallier 2015, 937).

section as the culmination of Rawls's reasoning on stability, where all the elements of his theory of the good and his moral psychology are combined and tied together.[10]

Their view has definitely some textual support. Rawls presents the problem of stability in game-theoretic terms in two sections of *TJ*. In the section "Economic Systems" he presents two problems that arise in distributive systems, the problems of isolation and assurance. He compares the problem of isolation with "the general case of the prisoner's dilemma of which Hobbes's state of nature is the classical example", which arises "whenever the outcome of the many individuals decisions made in isolation is worse for everyone than some other course of action, even though, taking the conduct of the others as given, each person's decision is perfectly rational" (Rawls 1971, 269). The assurance problem in contrast is associated with the fact that "each person's willingness to contribute is contingent upon the contribution of the others" (Rawls 1971, 270).[11]

In another section of *TJ* Rawls writes that there are two tendencies leading to instability. The first one, arises from the fact that some citizens may be tempted to avoid fulfilling their duties, if they believe that they will still benefit from the distribution of public goods. The second one, arises from the fact that citizens may stop contributing their share if they believe or suspect that others are not contributing theirs. Rawls notes that this instability is likely to be strong, when the risk of sticking to the rules when others are not, is too high (1971, 336).

There is also textual support that the congruence argument aims to solve the free-rider and mutual assurance problems. In the end of the section "The Good of the Sense of Justice" Rawls claims that "the hazards of the generalized prisoner's dilemma are removed by the match between the right and the good" (1971, 577). The hazards of the generalized prisoner's dilemma refer to the instability that threatens a well-ordered society by the forces that prompt its citizens to skip their duties. As I mentioned above, Rawls identifies these forces as the temptation to adopt a free-riding behavior and the lack of assurance that others will comply with their duties. The match between the right and the good, namely their congruence, restricts the temptation to act as a free-rider because citizens in a well-ordered society would judge that they are better off by being just persons than

[10] This claim is supported by the fact that at the beginning of the section Rawls writes that "now that all the parts of the theory of justice are before us, the argument for congruence can be completed" (Rawls 1971, 567). Freeman (2003, 277) for example writes that the congruence argument "begins in Part III of *Theory of Justice (TJ)*, is developed for over 200 pages, and culminates (in Section 86) at the end of a very long book".

[11] It is important to note that Rawls makes clear that his view on stability is different from that of Hobbes. Rawls aims to ensure a moral stability or a "stability for the right reasons" and not a mere modus vivendi. Unlike Hobbes, the aim is not to achieve the obedience to law by the existence of some external mechanism that delivers sanctions. Instead, Rawls purports to show that the two principles of justice can be internalized by the citizens of a well-ordered society and become a part of their good.

they would be if they took advantage of others. And the fact that it is publicly known that everyone is motivated in the same way,[12] solves also the problem of mutual assurance.[13]

II. THE SPECIAL PSYCHOLOGIES AND THE HAZARDS OF ENVY

The game-theoretic view seems well justified if one takes into account the excerpts mentioned above. Yet there are other parts of this very long book that do not confirm this view and raise doubts as to whether Rawls was focused solely on the free-rider and mutual assurance problems. A large part of the chapter "The Good of Justice" seems to be completely independent of the overall game-theoretic thinking. More specifically, in the three sections where he discusses the problem of envy (80-82), Rawls seems to regard envy as a separate source of instability that threatens a well-ordered society. There are a number of passages that support this claim.[14]

In the section "The Problem of Envy" Rawls presents envy as a collectively disadvantageous propensity which is dangerous and worsens the situation of all the parties involved. He writes that after a conception of justice is selected in the original position, we must check if it is going to arouse envy at such an extent that "the social system becomes unworkable and incompatible with the human good" (Rawls 1971, 531). He then notes that regarding justice as fairness, we have to assess if the inequality allowed by the difference principle is so acute that it generates destructive feelings of envy (Rawls 1971, 532). Consequently, Rawls explains the reasons why envy is so dangerous. First, when we envy other persons we no longer value what we have and this loss arouses hostile feelings. Second, we are willing to deprive them of their benefits even if we have to

[12] The public knowledge of this fact could mean for example the data published by tax authorities and other public institutions.

[13] This solution to the assurance problem is briefly mentioned by Rawls (1971, 336) in another part of *TJ*, where he says that the assurance problem "is to maintain stability by removing temptations of the first kind, and since this is done by public institutions, those of the second kind also disappear, at least in a well-ordered society".

[14] Rawls uses some aspects of game theory to clarify the special psychology of envy. He notes for example that envy can be collectively disadvantageous. "When others are aware of our envy, they may become jealous of their better circumstances and anxious to take precautions against the hostile acts to which our envy makes us prone. So understood envy is collectively disadvantageous: the individual who envies another is prepared to do things that make them both worse off, if only the discrepancy between them is sufficiently reduced". Rawls (1971, 532). At the same time envy can also prompt us to excel, which can be collectively advantageous. "A somewhat different case is that of emulative envy which leads us to try to achieve what others have. The sight of their greater good moves us to strive in socially beneficial ways for similar things for ourselves" (1971, 533). In *TJ* Rawls examines how the first kind of envy can threaten the stability of justice as fairness. This threat is triggered by psychological and social conditions that are separate from the ones that are related with the problems of mutual assurance and free-riding behavior. The problem with envy is not that we don't have the assurance that others are doing their share, is that we judge ourselves "happy or unhappy only by comparison with others" (Kant 2009, 27). I would like to thank an anonymous referee for comments regarding the relation between envy and game theory in *TJ*.

give up something ourselves. Third, when others know that we envy them they may take preventive measures against possible hostilities (Rawls 1971, 532).

The problem of envy is mentioned by Rawls also in the chapter "The Original Position", where he presents stability as one of the criteria that the parties have to take into account when they choose a conception of justice. In this chapter Rawls relates the problem of envy with that of stability, while he doesn't make any reference to the mutual assurance problem. In particular, he notes that in the last part of *TJ* he will try to show that the selected principles lead to a well-ordered society where envy and other destructive feelings are not likely to be strong and that the conception of justice undermines the circumstances that trigger disruptive attitudes (Rawls 1971, 144).

The above passages from *TJ* indicate that Rawls treated the problem of envy as a distinct destabilizing power which is not part of a collective action problem and a prisoner's dilemma. They also indicate that Rawls regarded the problem of envy as an important test for the two principles of justice.

One possible reply by the supporters of the game-theoretic view is that the examination of the problem of envy is not distinct from the overall reasoning on the assurance and isolation problems and the congruence between the right and the good. For example, Weithman (2010, 141) believes that the conclusions Rawls draws about the mitigation of envy by the two principles of justice are integrated in the final congruence argument. He thinks that these conclusions provide an important supplement on Rawls's argument that citizens of a well-ordered society regulated by the two principles of justice would have a desire to avoid the psychological cost of hypocrisy and the desire to have ties of friendship. These two desires, along with the desire to express themselves as free and equal beings and the desire to participate in forms of life that call forth their own and others talents, form the grounds of congruence between the right and the good.[15]

Regarding the first desire, Weithman (2010, 141) points out that it is because of "the argument in 82" that a typical member of a well-ordered society "would regard the costs of hypocrisy and deception as high relative to the benefits of the wealth she could get above her fair share".[16] What is the "argument in 82" according to Weithman? It is the argument that members of a well-ordered society would not feel envy for those who have more wealth, because their status is not defined by their relative position in the distribution of wealth but by the position of equal citizenship. If they know that others respect them as equal citizens, they would not be moved to seek wealth and economic status as a means to self-respect. So, the fundamental assertion in Rawls's argument is

[15] The final congruence argument in *TJ* is that persons who have these four desires would also have a desire to act justly, because by doing so they promote their good.

[16] Weithman argues that the idea that members of a well-ordered society would regard as very high the cost of hypocrisy is not obvious and it needs to be explained. For Weithman, when Rawls says that the cost of hypocrisy would be higher in a well-ordered society he does not mean it in absolute terms. Instead, he implies that they would be higher relative to what citizens could get by paying them, for example, the greater wealth they would enjoy by cheating on their taxes.

that "the position of equal citizenship answers to the need we might have thought people had for economic status" (Weithman 2010, 142). The same holds for the second desire, the desire to have ties of friendship. The members of a well-ordered society would judge that the potential costs to those with whom they have ties of friendship give them strong reasons to treat their sense of justice as supremely regulative, since they don't need more wealth for self-respect.

Freeman (2007, 153) believes that the argument from the absence of envy is just one part of the "peculiar array of arguments in chapter 9 of *Theory*" that shows that the sense of justice is not "in many respects irrational and injurious to our good".[17] In particular, he maintains that in the final chapter of *TJ,* Rawls argues that the sense of justice is not arbitrary, entirely conventional and self-destructive, is not grounded in a self-debasing submission to authority, it accounts for the good of community and it doesn't mask a lack of self-worth and a sense of failure and weakness. This array of arguments, in which the absence from envy is included, is presented by Freeman as part of Rawls's main argument for the good of justice, the congruence argument.

Even if Weithman and Freeman were right[18] and the absence of envy plays a role in the congruence argument, this would not be a proof that Rawls treats the problem of envy as an intrinsic part of a collective action problem. There are numerous passages in *TJ* where Rawls presents envy as a serious threat to the stability of a well-ordered society, a threat which is distinct from the hazards of a generalized prisoner's dilemma and calls for a distinct solution. The destructive character of envy introduces a difficulty that goes beyond the generalized prisoner's dilemma. It is not only that we don't have the assurance that others are doing their share, but it is also the sidelong glance we cast at one another, taking an interest in others relative position in the distribution of wealth, income and social status.[19]

In the sections 80 to 82 Rawls demonstrates why the two principles of justice alleviate this threat, at least more than the principle of utility. The congruence argument on the other hand does not deal with the disruptive effects of envy. The argument that the sense of justice is compatible with a person's good can be an answer to the problems of mutual assurance and isolation but is irrelevant to the problem of special psychologies.

[17] Freeman responds to Brian Barry's claim (1995) that the acquirement of a sense of justice is sufficient to prove stability within Rawls's framework.

[18] I believe that the absence of envy is one of the reasons that citizens see their political society as a good, but I don't think that it functions as a "supplement" argument in the way presented by Weithman. My claim that we have to treat the absence of envy as a distinct argument for the stability of a well-ordered society doesn't mean that it doesn't have a supportive role in the congruence argument. The mitigation of envious feelings contributes to the good of the citizens of a well-ordered society. But this shouldn't overshadow the fact that Rawls treats envy as a distinct source of instability which is inextricably linked with distributive justice.

[19] This point is stressed by William A. Edmundson (2017, 105) who claims that envy and the special psychologies introduce a new kind of difficulty.

III. THE ARGUMENT FROM THE ABSENCE OF ENVY

I have argued that the game-theoretic view underestimates the importance of envy in Rawls's account of stability by not recognizing it as a distinct source of instability. I have also argued that the proponents of this view treat the congruence argument as the culmination of Rawls's overall argument for stability.

My thesis is that we can have a better understanding of Rawls's reasoning on stability if we see the congruence argument as only one part of a larger argument. Another important part is the argument from the absence of envy. So, on my interpretation, the overall stability argument is developed in three parts. In chapter 8, Rawls presents a moral psychology designed to show how people in a well-ordered society can acquire a sense of justice. In chapter 9, Rawls argues that a) members of a well-ordered society would not be moved be feelings of envy and b) that their sense of justice would be congruent with their good. The argument from the absence of envy aims to show how the two principles of justice mitigate the destabilizing feelings of envy. The congruence argument purports to remove the hazards of the generalized prisoner's dilemma.[20]

I will try to present a detailed reconstruction of the argument from the absence of envy and its connections with the bases of self-respect. I will emphasize how it supports the justification of the two principles of justice not only in comparison with the utilitarian principle but also with a more egalitarian conception. Consequently, I will look at the survival of this argument in *Justice as Fairness: A Restatement.*[21]

In *TJ* Rawls argues that self-respect is a primary good and "perhaps the most important" one (Rawls 1971, 396). This emphasis reveals the importance of self-respect in the choice of the two principles of justice from the original position. It is equally important though for the relative stability of justice as fairness as a conception of justice. In section 29 Rawls defines self-respect as the sense of one's own worth. It is the sense that the plan of life we have chosen is worth carrying out. He gives a more nuanced definition in the chapter "Goodness as Rationality". There he claims that self-respect or "self-esteem" has two aspects. The first aspect, is how we sense our own value, the conviction that our life plan is worth carrying out. The second one, is the confidence we have in our ability to execute this plan (Rawls 1971, 440).

[20] In chapter 9 Rawls (1971, 513) discusses "various desiderata of a well-ordered society and the ways in which its just arrangements contribute to the good of its members". He argues that justice as fairness demonstrates the objectivity of judgments of justice, something that can assure us that our sense of justice is not arbitrary and conventional. He shows also that justice as fairness is not grounded in a self-debasing submission to authority, that it can be combined with the good of community and that it enables citizens to express their nature as free and equal human beings. These are also considerations in favor of the stability of justice as fairness. The main difference is that the argument from the absence of envy and the congruence argument correspond to distinct sources of instability.

[21] Cited hereafter as *JFR*.

This is the reason that makes self-respect a primary good and perhaps the most important one. Without it "nothing may seem worth doing, or if some things have value for us, we lack the will to strive for them [...] and we sink into apathy and cynicism" (Rawls 1971, 440).

Rawls stresses that there are two circumstances that support the first aspect of self-esteem, the sense of our own worth. The first one is having a rational plan of life that satisfies the Aristotelian principle and the second one is finding that others appreciate who we are and what we do (Rawls 1971, 440). So there is a reciprocal feature on how we secure our self-respect. Our self-esteem is dependent on how we think our fellow citizens value us. Rawls's strategy is to show that the citizens of a well-ordered society regulated by the two principles of justice would be immune to destabilizing feelings such as envy because they would have a strong sense of self-respect.[22] In section 81, Rawls claims that the conception of justice as fairness supports the self-respect of citizens, because they all have the same basic rights and are treated as equals. This is based on the assumption that in a well-ordered society liberties and rights are more important for the self-esteem of citizens than their income share.[23] Rawls (1971, 545) gives two reasons why liberties and rights should have priority for securing the primary good of self-respect. First, if their self-respect was anchored in their wealth and income, then the pursuit of status and self-esteem would be a zero sum game, where the improvement of one person's position would lower that of someone else.[24] Second, it would be irrational to accept equal division of wealth to secure equal status and self-esteem, since there are compelling grounds for allowing an unequal distribution of wealth. Thus, it is more preferable to support the primary good of self-respect by giving priority to the basic liberties and defining the same status for all.

Rawls offers two additional points on why justice as fairness would secure self-respect. The first is that in a well-ordered society the less favored know that the greatest advantages others enjoy work also for their benefit. The difference principle permits deviations from strict equality only if this would make the less advantaged worst of. This reciprocal feature of the difference principle makes it easier for them to accept the disparities between themselves and others.[25] The second point is that justice as fairness does not associate the distribution of income and wealth with moral virtue and excellence. So the less fortunate

[22] Rawls makes a straightforward connection between envy and self-respect by assuming that the main psychological root of the liability to envy is a lack of confidence in the worth of our plans of life and our ability to execute them. By contrast, someone with a robust self-esteem has no desire to level down the advantages of others.

[23] "The basis for self-esteem in a just society is not then one's income share but the publicly affirmed distribution of fundamental rights and liberties. And this distribution being equal, everyone has a similar and secure status when they meet to conduct the common affairs of the wider society" (Rawls 1971, 544).

[24] This is another point where Rawls uses some aspects of game theory in his overall discussion of the problem of envy.

[25] It is a principle which expresses the ideas of fraternity and reciprocity, since citizens feel that the others care for their good and they don't try to take advantage of their bad luck.

have no reason to view themselves as inferior. In addition, Rawls (1971, 537) notes that given the background institutions of a well-ordered society, inequalities are not extreme and that the various social unions which flourish in a well-ordered society reduce the painful visibility of these inequalities, because citizens tend to compare their situation with others that hold positions similar to their aspirations. These features make it less possible for the less advantaged to experience humiliation.

For all the above reasons, Rawls concludes that justice as fairness is not challenged by outbreaks of envy and that it seems relatively stable. In the end of section 81 he states that so long "as the pattern of special psychologies elicited by society either supports its arrangements or can be reasonably accommodated by them, there is no need to reconsider the choice of a conception of justice" (Rawls 1971, 541). This could lead to the false conclusion that the argument from the absence of envy is applied by Rawls only to show that justice as fairness is a stable and feasible conception that would not be destabilized to a troublesome extent by destructive feelings of envy. There is no doubt that one of the main aims of Rawls is to demonstrate how justice as fairness removes the hazards of envy, in the same way that the congruence argument removes the hazards of the generalized prisoner's dilemma. But the absence of envy is also an argument for the relative stability of justice as fairness. In the third part of *TJ* and in parallel with other considerations regarding the problem on stability, Rawls compares the relative stability of justice as fairness with other conceptions of justice. This comparison serves as an important confirmation of the arguments in favor of the two principles of justice made in the first part of the book. As Rawls (1971, 498) notes, "other things equal, the preferred conception of justice is the most stable one".[26]

One of the reasons the two principles of justice are more stable from the utility principle or from a conception of strict equality, is that they lead to social structures in which envy is not likely to be strong (Rawls 1971, 144). In the third part of *TJ*, in the section "The problem of relative stability", Rawls (1971, 499) argues that in a well-ordered society regulated by the two principles of justice, citizens know that others don't take advantage of their bad luck and that they have an unconditional concern for their good and this caring strengthens their self-esteem. The utilitarian conception instead, is destructive of the self-esteem of those who rank low in

26] There is an ambiguity in the justificatory role of stability in *TJ*. There are passages where Rawls (1971, 498) claims that relative stability is crucial in the justification of the two principles: "There seems to be no doubt then that justice as fairness is a reasonably stable moral conception. But a decision in the original position depends on a comparison: other things equal, the preferred conception of justice is the most stable one. Ideally, we should compare the contract view with all its rivals in this respect, but as so often I shall only consider the principle of utility". But a few pages later Rawls (1971, 504) writes: "These remarks are not intended as justifying reasons for the contract view. The main grounds for the principles of justice have been presented. At this point we are simply checking whether the conception already adopted is a feasible one and not so unstable that some other choice might have been better".

the distribution of wealth and income. Given how crucial self-respect is for the psychological immunity against destructive feelings of envy, the two principles of justice are more stable than the principle of utility. The greater stability of the two principles is already implied in the first part of *TJ,* in the section "Main grounds for the two principles", where Rawls (1971, 178) claims that one of the reasons for the choice of the two principles of justice in the original position is that they give greater support to self-respect and this increases the effectiveness of social cooperation. If they chose the utility principle they wouldn't have the support that the reciprocal character of the two principles provides to their self-esteem.

The test of envy is also important for the comparison of the two principles of justice with a conception of strict equality. Although, such a conception is initially excluded in the original position, the parties would have reason to reconsider it if the inequalities allowed by the difference principle are so great that excite envious feelings to a dangerous extent. This could be the case for example "if how one is valued by others depends upon one's relative place in the distribution of income and wealth" or if wealth, income and other goods are "fixed and cannot be enlarged by cooperation" (Rawls 1971, 545). Rawls examines this possible scenario in the section "The Grounds for the Priority of Liberty" and he concludes that strict equality should not be considered as a solution, since in a well-ordered society the self-respect of citizens is secured by their equal civic status.[27]

IV. THE REVISED ACCOUNT OF STABILITY

The case that the third part of *TJ* contains an argument from the absence of envy, apart from the congruence argument, is supported by Rawls's account of stability in *JFR*.

In this book Rawls reformulates the presentation and defense of justice as fairness, integrating the changes he made in *PL* and distancing himself from the partially comprehensive view of the good that he presupposed in *TJ*. This reformulation shows how justice as fairness can be understood as a form of political liberalism and not as a comprehensive doctrine. A comprehensive doctrine is a set of beliefs that "covers the major religious, philosophical, and moral aspects of human life in a more or less consistent and coherent manner" (Rawls 2005, 58). It is comprehensive in the sense that it covers a wide range of values and is not restricted to political ideals. In *TJ* justice as fairness was presented as a comprehensive

[27] It is worth mentioning that Rawls acknowledges more than once that a socialist regime could be considered as a possible solution to the problem of stability. In his paper "Fairness to Goodness", he writes that "the principles of justice do not exclude certain forms of socialism and would in fact require them if the stability of a well-ordered society could be achieved in no other way" (Rawls 1975, 546). In *JFR*, Rawls raises the question if a liberal socialist regime would have a better chance of stably realizing justice as fairness than a property-owning democracy.

doctrine because it contained ideals of personal conduct.[28] Rawls recasts all the basic arguments and ideas of *TJ* and presents them in a way that is consistent with a political conception of justice. A political conception of justice is not derived from a single comprehensive doctrine but from ideas implicit in the public political culture of a democratic society and it is compatible with a wide range of different comprehensive doctrines.

One of the arguments he revises is of course the stability argument. In *PL* Rawls makes a radical change in his account of stability, introducing the idea of an overlapping consensus between different comprehensive doctrines and repudiating the idea of congruence between the right and the good. As expected, this change is preserved in *JFR*. Rawls (2001, 181) writes that one of the main questions regarding stability is whether in view of the general facts that characterize a democracy's political culture, and in particular the fact of reasonable pluralism, the political conception can be the focus of an overlapping consensus. The attempt to show congruence between the right and the good is abandoned, since it can't be assumed that citizens accept a particular comprehensive view to which justice as fairness belongs (Rawls 2001, 187). Along with the congruence argument, Rawls abandons in *JFR* the whole game-theoretic view. There is not any mention to the prisoner's dilemma or to any other kind of games.[29]

But although Rawls has removed the congruence argument in *JFR*, he still affirms that the two principles of justice should be tested against the destabilizing effects of envy and other special psychologies. In particular, Rawls (2001, 181) writes that the parties in the original position must check whether people who grow up in a well-ordered society acquire a strong sense of justice and are not moved to act by envy and spite. This is one of the two criteria for the stability of a conception of justice, along with the possibility of an overlapping consensus.[30]

If the test of envy was just a part of the congruence argument, one would expect that Rawls would stop using it after the revisions he made in *PL*. The fact that he

[28] These ideals are based on the Aristotelian principle and the Kantian conception of the person. In *TJ* Rawls assumes that citizens of well-ordered society would form their plans of life by giving priority to the cultivation of their talents and to their autonomy.

[29] There is a reference to the problem of mutual assurance, although it is not presented in a game-theoretic context and is viewed more as a necessary condition of stability and less as a distinct threat. Rawls (2001, 196) says that citizens are ready and willing to do their part in those arrangements provided they have sufficient assurance that others will also do theirs. This assurance is given by the existence of an overlapping consensus, since citizens know from past experience that others comply with liberal arrangements and come to have increasing trust in one another (Rawls 2001, 197).

[30] Rawls notes in *JFR* that the discussion in sections 81 and 82 of *TJ* addresses sufficiently the problem of envy and doesn't need substantial changes. The only addition to the argument from envy is the general remark that the more citizens see their political society as good and the greater their appreciation of the political conception in securing the three essentials of a stable regime, the less they will be prompted by envy and other special attitudes (Rawls 2001, 202).

maintains this test as a criterion of stability in *JFR*, which was published in 2001, supports the view that Rawls treats the destabilizing effects of envious feelings as a distinct problem from the hazards of a generalized prisoner's dilemma. In *TJ* the argument from the absence of envy solves the first problem and the congruence argument the second.

V. THE POLITICAL TURN

In *PL* Rawls abandons any attempt to prove that justice as fairness is a stable conception because it enables the congruence between the right and the good. He aims to show instead that justice as fairness can be the focus of an overlapping consensus between opposing conceptions of the good.

Despite of this sweeping change, the central elements of the account of stability presented in *TJ* remain intact in *PL*. Rawls still holds that we have a capacity for a sense of justice and that under the favorable conditions of a well-ordered society citizens will develop this capacity. Furthermore, he still views stability as a moral problem which is distinct from the Hobbesian concern for social order.

But does he still present the problems of assurance and isolation as possible sources of instability? In *Why Political Liberalism? On John Rawls's Political Turn* Weithman argues that in *PL* Rawls deploys the idea of overlapping consensus so that he can solve the mutual assurance problem and the prisoner's dilemma, this time taking into account the fact of reasonable pluralism. The problem now is that in a liberal democratic society there is a plurality of different comprehensive doctrines and each citizen needs an assurance that other citizens' conception of the good doesn't provide sufficient reason to act against the demands of justice. The existence and public knowledge of an overlapping consensus solves according to Weithman (2010, 297) this mutual assurance problem.

In *PL* however there is not a single mention on the prisoner's dilemma or any other collective action problem. One would expect that if Rawls had in *PL* a game-theoretic approach to stability he would have presented the nature of the problem he tries to solve in a game-theoretic way.

Rawls mentions though something else in the introduction of the second edition of *PL* which sheds more light on the problem of stability he tries to solve in *PL*. He notes that *TJ* and *PL* deal with two different sources of conflict (Rawls 2005, lviii):

> There are three main kinds of conflicts: those deriving from citizens' conflicting comprehensive doctrines; those from their different status, class position, and occupation, or from their ethnicity, gender and race; and finally, those resulting from the burdens of judgments. Political Liberalism mitigates but cannot eliminate the first kind of conflict, since comprehensive doctrines are politically speaking, unreconcilable and remain inconsistent with one another. However, the principles of justice of a reasonably just constitutional regime can reconcile us to the second kind of conflict. For once we accept principles of justice or recognize them as at least reasonable (even though not as the most reasonable) and know that our political and

> social institutions conform to them, the second kind of conflict need no longer arise, or arise so forcefully. I believe that these sources of conflict can be largely removed by a reasonably just constitutional regime whose principles of political justice satisfy the criterion of reciprocity. PL does not take up these conflicts, leaving them to be settled by justice as fairness (as in Theory) or by some other reasonable conception of justice. Conflicts arising from the burdens of judgments always, however, remain and limit the extent of possible agreement.

It is my contention that we can have a better understanding of the political turn if we see it as an attempt to address a source of conflict that didn't concern Rawls in the third part of *TJ*. In *PL*, Rawls doesn't mention income inequalities, the alienation of less privileged citizens and the problems that arise from the division of labor, not because he doesn't believe that they pose a threat to the stability of a well-ordered society, but because he focuses on the conflicts that arise from inconsistent and unreconcilable comprehensive doctrines and especially between religious and liberal comprehensive doctrines. This is the source of instability he is concerned with in *PL*. The extensive revisions he made to *TJ* serve the purpose of presenting a liberal conception of justice which can address this source of conflict. If we see these revisions as a more realistic attempt to solve the generalized prisoner's dilemma presented in *TJ* we don't grasp that *PL* marks a shift of focus in Rawls's project.

It is of course possible to describe the conflicts between religious and liberal comprehensive doctrines in game-theoretic terms. In a society with a plurality of comprehensive doctrines citizens can be uncertain as to whether one doctrine may seek to achieve national preeminence. They need an assurance that others are going to respect the demands of justice. If not, the mistrust between the different comprehensive doctrines might weaken citizen's commitment to their sense of justice. But this description fails to capture the real character of the political turn. It undermines the change of focus in Rawls's thinking. In *TJ* the problem of reasonable pluralism is ignored and there is not any discussion on how the values of a liberal democratic society can be compatible with the values of religious comprehensive doctrines. If we describe the conflicts between different comprehensive doctrines as a special case of a collective action problem, then we risk trivializing their depth.

VI. CONCLUSION

A fundamental question that Rawls addresses in *TJ* is that of political and social stability. Institutions must not only be just but they must generate their own support over time. If a just conception of justice fails to achieve this, then another one must be considered.

Rawls's argument for the stability of justice as fairness is developed in two stages. First, he aims to show that a well-ordered society regulated by the two principles of justice would bring about in its members an effective sense of justice. Second, he claims that their sense of justice would be strong enough to outweigh propensities to act otherwise.

This claim is based on the idea of congruence. The members of a just society would have a strong sense of justice because they would affirm that it is part of their good.

According to a game-theoretic view of Rawls's account of stability the congruence argument is the culmination of his overall argument for the stability of justice as fairness. This view is based on the assumption that Rawls's aim in the third part of *TJ* is to present a solution to a generalized prisoner's dilemma and the related problem of assurance. It finds textual support in section 86 of *TJ* where Rawls writes that the hazards of the generalized prisoner's dilemma are removed by the match between the right and the good. But other passages from *TJ* indicate that Rawls treats envy as a distinct source of instability which is not part of a collective action problem. In the sections 80-82 of *TJ* he shows that the two principles of justice would not generate feelings of envy to a socially dangerous extent.

My thesis is that the overall stability argument in *TJ* is developed in three parts. In the first part Rawls claims that people in a well-ordered society would acquire a sense of justice. He then argues that members of a well-ordered society would not be moved by feelings of envy and that their sense of justice would be congruent with their good. The last two parts of the argument respond to two different tendencies that might prompt citizens to act against their sense of justice.

This reading of the problem of stability in *TJ* can deepen our understanding of Rawls's political turn and is aligned with the revised account of stability in *JFR*.

manolatosalex@yahoo.gr

REFERENCES

Barry, Brian. 1995. John Rawls and the Search for Stability. *Ethics* 105 (4): 874-915.

Edmundson, William. A. 2017. *John Rawls: Reticent Socialist.* Cambridge: Cambridge University Press.

Freeman, Samuel. 2007. *Justice and the Social Contract. Essays on Rawlsian Political Philosophy.* New York: Oxford University Press.

———. 2003. Congruence and the Good of Justice. In *Cambridge Companion to Rawls,* edited by Samuel Freeman, 277-316. New York: Cambridge University Press.

Galisanka, Andrius. 2017. Just Society as a Fair Game: John Rawls and Game Theory in the 1950s. *Journal of the History of Ideas* 78 (2): 299-308.

Kant, Immanuel. *Religion within the Bounds of Bare Reason.* Translated by Werner S. Pluhar. Indianapolis: Hackett Publishing Co, 2009.

McClennen, Edward F. 1989. Stability and the sense of Justice. *Philosophy & Public Affairs* 18 (1): 3-30.

Quong, Jonathan. 2014. On the Idea of Public Reason. In *A Companion to Rawls,* edited by David A. Reidy, and Jon Mandle, 265-81. Chichester: Wiley Blackwell**.**

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

———. 1975. Fairness to Goodness. *The Philosophical Review* 84 (4): 536-54.

———. 2005 [1993]. *Political Liberalism.* New York: Columbia University Press.

———. 2001. *Justice as Fairness: A Restatement.* Cambridge MA: Harvard University Press.

Thrasher and Vallier. 2015. The Fragility of Consensus: Public Reason, Diversity and Stability. *European Journal of Philosophy* 23 (4): 933-54.

Weithman, Paul. 2010. *Why Political Liberalism? On John Rawls's Political Turn.* New York: Oxford University Press.