Nordic Journal of Political Economy

Volume 23 1996 Pages 3-23

Transmutation and Misrepresentation

Jon Elster

Jon Elster *

Transmutation and Misrepresentation

In a society with progressive taxation, those with high incomes have a strong interest in lower taxes. In arguing for tax cuts, however, they cannot simply appeal to their interest. They cannot say, "Congress should cut taxes because that's good for me". Instead, they will have to argue for their proposal in terms of the general good. By appealing to trickle-down effects and supply-side considerations they can claim that *everybody* will be better off if the rich get a tax break. If they make this argument in public, repeatedly, they may end up believing it themselves. Most people do not like to think of themselves as liars or cynics. To say one thing and think another is a source of tension and discomfort that can be removed by aligning one's thought on one's utterances. In fact, that tension need not even arise. Most people do not like to think of themselves as motivated only by self-interest. They will, therefore, gravitate spontaneously towards a world-view that suggests a coincidence between their special interest and the public interest.

This example suggests two mechanisms of more general application. First, people may have an incentive to *misrepresent* their motivations to others. This is the topic of the last section below. Second, they may be subject to psychic pressures that cause an original motivation to be *transmuted* into another. This is the topic of the following section.

I shall discuss three motives that can serve as inputs and engines of transformation: reason, interest, and passion. My conception of *reason* relies on work by Habermas. He argues that an agent who aims at *understanding* rather than *success* is committed to three "validity claims": propositional truth, normative rightness, and sincerity or truthfulness. He must be open to rational argument and willing to change his view as the result of such argument. It follows that a speaker who wants to *appear* – to himself or others – as aiming at understanding must also appear to be committed to these claims. It is this secondary or derived concern that is my main topic here. I shall not try to explore in any detail the conceptual implications of the claims; rather, I shall consider what agents must do to appear

* Colombia University, New York. This article is a condensed version of a chapter in my forthcoming book *Alchemies of the Mind*. Several issues that I had to neglect here for reasons of space are discussed there, notably the views of Festinger and Freud.

to be committed to them and what happens if this appearance breaks down.

The first claim is straightforward: in any communication about factual issues the parties share the assumption that there is *a fact of the matter* by virtue of which what they say is either true or false. Below I argue that by virtue of its objectivity truth can serve as a strategic resource: by restating a threat as a warning, a speaker may present an interest-based claim in the language of reason. The second claim is more controversial. Rather than identifying the normative variety of reason with a particular conception of justice, I shall count any impartial, disinterested and dispassionate motive as a reasonable one. In order to be viewed by oneself or others as having a morally or socially acceptable motivation, what matters is to be seen as moved by *some* impartial conception, not by any particular one. The third claim imposes, as we shall see, constraints of how the second claim can be met. An impartial conception that corresponds too closely to the speaker's interest, for instance, may not be seen as fully sincere.

Let me say a bit more about normative impartiality, which is the most important aspect of reason to be discussed below. Impartiality as such is not a conception of justice, but a necessary feature of any view that wants to be taken seriously as a conception of justice. It is a constraint on justice, not itself a conception of justice. Utilitarianism, for instance, is impartial in its insistence that in the calculus of welfare "each is to count for one and nobody for more than one". Rights-based theories are impartial to the extent that rights are assigned universally rather than selectively. Equal-distribution theories are impartial, as are theories advocating distribution according to need, merit or contribution. John Rawls' theory of justice is impartial by construction, being presented as the theory that rational individuals would choose behind a veil of ignorance.

For my purposes, two closely related features of impartial justice will prove especially important. First, as just noted, there are *many* conceptions of justice that satisfy the constraint of impartiality. As a consequence, claims that are motivated by interest or emotion will often be able to find an *impartial equivalent*, i.e. an impartial argument converging to the same conclusion. Second, impartial intent does not imply impartial effect. The law, in many contexts, allows rules that have "disparate effect" on different categories of individuals, while prohibiting those that embody "disparate intent". Thus a court might strike down a layoff rule that explicitly favors men over women or white over blacks, while allowing rules based on seniority which, because of the more recent entry of women and blacks into the labor force, have the same effect. If one could prove that seniority was adopted in order to produce that effect, it would be struck down. Yet because it is hard to prove intention *and* because seniority is prima facie fair, courts tend to respect the principle.

Impartiality, being both disinterested and dispassionate, has two antonyms: interested dispassionate behavior and passionate disinterested behavior. By *interest* I mean any motive, common to the members of some proper subgroup of society, that aims at improving the situation of that subgroup in some respect such as pleasure, wealth, fame, status or power. Subgroups made up of one individual form an important special case. For larger subgroups, interest will not be causally efficacious unless individual group members are motivated to embrace it. A politician may identify with his party because if he doesn't he won't get renominated. A worker may identify with his fellow workers because he views their welfare as part of his. A doctor may identify with his patients because of his professional norms. Interest by defini-

tion is partial, as it does not extend beyond the subgroup. I shall assume that in the pursuit of their interest people follow the canons of instrumental rationality. By *passion* I mean any kind of emotion that is characterized by arousal and valence to some non-negligible extent. By this loose characterization I only intend to exclude the aesthetic emotions and other low-level emotional states such as mild boredom or mild embarrassment. This terminology is a mere rhetorical device, with no substantive implications.

In the following section I discuss how the mind is capable of transmuting

- interest into reason
- passion into reason
- interest into passion
- passion into interest
- passion into passion

In the last section I discuss how we misrepresent

- interest as reason
- interest as passion
- passion as reason
- passion as passion
- passion as interest
- reason as interest
- reason as reason

The motivation behind these transformations can also be one of the three motivations themselves. In transmutation, the motivating force is a negative emotion triggered by awareness of the input motivation. In misrepresentation, the motivating force can be either of the three motivations.

As a final preliminary comment, note that transmutation and misrepresentation, as I define them, are more specific than the more general phenomena of lying to oneself or to others. Most cases of self-deception involve changes of beliefs about the world, not changes in beliefs about one's own motivation. The paradigm case is "I want X to be the case; therefore I believe X to be the case", rather than "I want to desire X out of motive Y; therefore I do desire X out of motive Y". Similarly, most cases of deception involve professing beliefs about the world that one does not hold rather than motivations one does not hold. The paradigm case is "If I profess belief X others will punish me, therefore I will profess belief Y", rather than "If I profess motivation X for desiring Y others will punish me, therefore I will profess motivation Z for desiring Y". Transmutation and misrepresentation are motivated by the desire to be or to be seen as a *certain kind of person*, who is incapable of holding and acting upon certain kind of motivation, not by the desire to achieve pleasure or avoid pain. We shall see, however, that in these transformations what changes is not simply the motivation, but also what one is motivated to do. They are, as it were doubly deceptive: motivation X for desiring Y is transformed into motivation Z for desiring W. The reason why Y is transformed into W is found in the *constraints* on transformation that I discuss extensively below.

Transmutation

To illustrate the processes of transmutation I draw on some of my earlier studies of individual and collective decision-making. These include the behavior of parents in child custody disputes (Elster, 1989a, Ch.III), collective wage bargaining (Elster, 1989b, Ch.III), the allocation by institutions of scarce resources and necessary burdens (Elster, 1992, 1995a), and the adoption of new political constitutions from the Federal Convention to the present (Elster, 1993, 1994, 1995b, 1995c, 1995d, 1995e). In all these situations, we can identify the voices of reason, interest, and passion. In a child custody dispute, for instance, the mother may be motivated by

her concern for the child's welfare (reason), by her desire to get custody (interest), or by her animus against the father (passion). In constitution-making, we may identify some elements as depositories of reason (the juridical system), others as bearers of interest (political parties) and still others as motivated by ethnic or religious passions. In addition to these studies, I take examples from laboratory experiments, novels and plays. Sometimes, I also use the last-resort procedure of invented examples. These are *not* be taken as confirming illustrations of the more abstract propositions, only as a way of making them more vivid.

Interest into reason. Instances of "self-serving conceptions of impartiality" are numerous. In collective bargaining, for instance, the use of self-serving conceptions of fairness is very common. The choice of a reference group compared to which one's wages can be argued to be unfairly low, for instance, is highly subject to self-serving manipulations (Elster, 1989b, Ch.VI). Because everything is a little bit like everything else and because intuitions about fairness are sensitive to a great variety of factors, one can usually identify a better-paid group that is similar in some normatively relevant respect. Thus in wage bargaining between a teachers' union and a school board, the union would tend to view neighboring high wage districts, and the board to view low wage districts, as comparable (Babcock and Olson, 1992, Babcock, Wand and Loewenstein, 1996).

In "local justice" the use of seniority as a criterion for layoffs is perhaps the most clear-cut case of self-serving conceptions of fairness. When unions advocate this principle, they often refer to ideas of desert and merit. As the senior workers have "given the best years of their life" to the firm, it is only fair that they should be retained over new hires. The argument is weak (what else should they have done with their life?) and quite likely a mere dressing-up for self-interest. As long as the workers have reason to believe that the firm will never lay off more than half of them, the workers with the greatest seniority will vote for the principle out of self-interest. Yet this motivation is very likely to be transmuted into a genuine belief that seniority is more fair. There is also a widespread belief, and not only among the elderly, that old people should be given priority in various contexts, because of "what they have done for society". The idea that entitlements are generated by the mere passage of time has wide and strong appeal.[1] When it coincides and fuses with self-interest, it may be irresistible.

In constitution-making processes, one regularly observes that the interests of different groups are defended in impartial language. At the Federal Convention, this tendency was strikingly illustrated in the debates between the small and the large states over their respective representation in the Senate. The small states systematically argued for equality of representation, the large states for proportionality. Although this confrontation involved some threats of force, impartially phrased argument had a more central place. Both sides, in fact, were able to defend their views by appeals to fairness and justice. There were obvious arguments from equality – equal representation of the states versus equal representation of individuals.[2] In other constitution-making contexts, we find some

1. See Zajac (1995, p.121-22) who cites as a "striking example of status quo property or equity rights [..] the attempt to 'vintage' utility rates, that is, to charge 'old' customers a different, usually lower rate than 'new' ones".
2. See for instance Madison in Farrand, ed., (1966, vol. I, p.151) versus Dickinson (*ibid.*, vol. I, p.159).

actors dressing up their interests in consequentialist garbs, while others appeal to non-consequentialist values. In the recent transitions to democracy in Eastern Europe, for instance, most parties favored the electoral laws that favored them. Small parties favored proportionality (with low thresholds), large parties favored majority elections in single-member districts (or proportionality with high thresholds). The former tended to defend their views in terms of democratic values, the latter in terms of the need for an efficient and stable government. Similar self-serving biases were observed in the design of the presidency. If a group has a strong candidate for this position, its interest is to have a strong presidency and that of the opposition to have a weak one. The corresponding impartial arguments tend to be, for the former, that the difficult period of transition to democracy requires a strong leadership and, for the latter, that a strong man at the head of the state might re-create an authoritarian regime.

There is also experimental evidence that people tend to choose conceptions of fairness that favor themselves. Linda Babcock, George Loewenstein and their collaborators have carried out laboratory experiments in which subjects are assigned to the role either of plaintiff or of defendant in a tort case and asked to negotiate a settlement (Loewenstein et al., 1993, Babcock et al., 1995). They were also asked to predict the award of the judge and to assess what they consider a fair out-of-court settlement for the plaintiff. They found that plaintiffs predicted higher awards and assessed higher fair-settlement amounts than defendants, and that pairs of subjects who reached more similar predictions and assessments were more likely to settle than those who reached very different assessments. The first finding is clear evidence of a self-serving bias in conceptions of fairness. The second finding suggests (but does not prove: see below) that the joint effect of self-serving biases works against the interest of the parties. Let me pursue that idea for a moment.

Paul Veyne suggests, as a general maxim, that beliefs born of passion serve passion badly (Veyne, 1976). Similarly, the second finding just cited might seem to show that beliefs born of interest serve interest badly. Yet the *probability* of settlement is only one of the factors that affect the expected outcome of the parties. The *amount* of settlement, in the cases where an agreement is reached, also matters. If a biased perception of fairness reduces the probability, it also tends to increase the amount. "For every dollar's increase in the defendant's perceived fair-settlement value, actual settlements rose, on average, by 50 cents", or, to put it the other way around, when the parties settled the defendant paid less the greater his bias.[3] As the authors note, "these benefits of a sense of entitlement have to be weighed against the increased risk of not settling". In general, the net effect is indeterminate.[4] Beliefs born of interest may serve interest – or not.

I conclude this discussion of the transmutation of interest into reason by introducing an idea that will prove important at several

3. Loewenstein et al. (1993, p.152-53). They add that "Curiously, however, plaintiff's predictions and fairness values did not have a significant effect on settlement values."
4. Elster (1989b, Ch.II) discusses cases in which pre-bargaining behaviors of the parties have a negative impact on the probability of agreement and a positive impact on the amount agreed upon, as well as cases in which they increase the share of the total while reducing the size of the total to be shared. These behaviors are, however, intentionally chosen and would presumably not be undertaken unless the expected net effect was positive. The fairness bias, although shaped by interest, is not similarly guided by interest.

places below. Even assuming a self-serving bias – a tendency that is not present in everybody, and may in some be replaced by the opposite, self-denying bias – we should not assume that it will induce in all situations the conception of impartiality that is optimal from the point of view of self-interest. The selection of a conception of fairness is subject to *constraints* that prevent us from picking and choosing conceptions of fairness à la carte according to what best serves our interest. I shall consider two such constraints, which I shall refer to as the *consistency constraint* and the *imperfection constraint.* While the consistency constraint certainly applies to transmutation as well as to misrepresentation and the imperfection constraint no less certainly applies to misrepresentation, I am less sure about the role of the imperfection constraint in transmutation.

The consistency constraint arises because the impartial conception adopted on a specific occasion has to be consistent with impartial conceptions adopted on earlier occasions. If it is not – if the agent opportunistically adjusts his idea of impartiality to what serves his interest on any given occasion – it will be psychologically difficult for him to maintain the belief that he is not motivated by self-interest. Some people may be capable of this feat of self-deception, but most are not. An impartial conception, once adopted, is perceived as binding and objectively valid in a way that constrains frictionless adjustment to new situations. The same need for self-respect that causes us to justify self-interested behavior by impartial considerations may also prevent us from changing our conception of impartiality when it no longer works in our favor. In character development, therefore, much would seem to depend on the accidental order in which we are exposed to circumstances in which judgments of fairness or other impartial arguments might arise. By the combination of self-serving bias and the need for consistency people may get locked into conceptions of impartiality that bear no recognizable relation to their overall interest later in life.

The work by Babcock, Loewenstein and co-workers cited above also offers evidence that consistency serves as a constraint on self-serving biases. The findings cited above do not by themselves show that failure to reach agreement is *caused* by the self-serving conceptions of fairness with which they are correlated. "Perhaps an unmeasured factor, such as variation in the character trait of the negotiators, caused the same people to exhibit the self-serving bias to negotiate in a manner than impeded settlement." (Babcock et al., 1995, p.1338). To test this possibility, they ran a variant of the experiment in which the subjects had to make their assessments of fairness *before* they knew whether they were going to take the role of the plaintiff or the defendant in the negotiation process. They found that "there were four times as many disagreements when bargainers knew their roles initially than when they did not know their roles". (Babcock et al., 1995, p.1339). The idea of a consistency constraint seems to make sense of this result: once the bargainers have decided behind "the veil of ignorance" what would be a fair settlement, they are stuck with that assessment and less likely to make opportunistic adjustments when they find our where their interest lies. Matthew Rabin refers to this phenomenon as *moral priming* (Rabin, 1995).

Conceptions of impartiality are not entirely irreversible. They may change under the pressure of changing interest, but the impact is often lagged. In the 1930s, wages of Swedish metal workers were below those of construction workers. The strong dissatisfaction of the metal workers with the existing wage differentials was a major cause of the

move towards centralized bargaining with its greater emphasis on inter-industry wage equality (Swenson 1988, pp.43-53). Later, when the metal workers became the high-wage outliers, they were bound by their past appeals to solidarity. As early as "at the beginning of the war, they were already some who thought that Metall had blundered by becoming the standard bearer for the idea of solidaristic wage policy in 1936. Certainly, it had been the underdog then, but now that they were better off, it gave them the moral obligation to show solidarity even when it was to their disadvantage." It took fifty years for the norm of equality to lose its grip on the metal-workers' union.[5]

The imperfection constraint arises because a *perfect* coincidence between self-interest and the impartial argument would often be too transparently opportunistic. To be credible to oneself or to others, the impartial argument has to deviate somewhat from the policy that, if adopted, would promote one's interest maximally. It should not deviate too much, of course, because then it might not promote one's interest at all. The optimal policy, therefore, has to strike a balance between interest and the appearance of disinterestedness. This constraint is certainly important in misrepresentation of interest, as I try to show below. I suspect that it is also important in transmutation. In the local-justice cases that I cite to illustrate the imperfection constraint on misrepresentation, for instance, transmutation may also be at work. Yet direct evidence for the imperfection constraint in transmutation is hard to come by.[6]

Passion into reason. Unavowable emotions may be transmuted and rendered acceptable by rewriting the script. There is a tendency, for instance, for irrational anger to rewrite the script so as to justify the emotion felt. In *On Anger*, Seneca says, "Reason wishes the decision that it gives to be just; anger wishes to have the decision which it has given seem the just decision." He also writes: "Men whose spirit has grown arrogant from the great favor of fortune have this most serious fault – those whom they have injured they also hate (*quos laeserunt et oderunt*)". Similarly, the shame induced by the painful emotion of envy may cause one to redefine the situation in a way that allows one to experience instead the wonderfully heady emotion of righteous indignation.

Prejudice, too, is a form of passion that can be dressed up as reason. In a study of attitudes towards blacks and affirmative action, Paul Sniderman and co-authors discuss two views of the relation between passion, reason and policy preference in the area of affirmative action. (Sniderman, Brody and Tetlock, 1991, Ch.6). The commonsense view is that people who *either* have a favorable (unfavorable) emotional attitude towards black *or* believe that the problems of blacks are due to factors outside (within) their control, are induced to adopt policy preferences in favor of (against) affirmative action. This view, however, is not confirmed empirically. Rather the data suggest that people go from their emotional attitude towards blacks directly to the policy conclusion, and then go backward to adopt the view of internal vs external control

5. Elster (1989b) p.160 ff. . For other examples of the role of consistency arguments in wage bargaining, see *ibid.*, pp.239-41.
6. It should be possible to test the hypothesis experimentally, by offering subjects the choice between several maxims of fairness and see whether they choose one that corresponds moderately to their interest or one that offers a perfect fit. To induce transmutation rather than misrepresentation, one could tell them that actual allocations would be made by applying an average of the fairness proposals made by different subjects, and (non-veridically) that no one would know who chose which maxim.

that justifies the preferred policy. "I don't like blacks; therefore, they should not receive assistance; therefore, they do not deserve assistance." Reason is the handmaid of passion.

Interest into passion. The transmutation of an interest in some object A into a passion for A may be illustrated by the phenomenon of marrying for money. Because there is something incongruous and shameful about this procedure, at least in modern societies, those who engage in it will sometimes find a way to persuade themselves that they are really marrying for love. The love thus generated will fall short of Stendhal's "amour-passion" which is characterized by reckless disregard of interest, and coincide more closely with his "amour-goût", which never goes against interest. Yet for all that it may still be passion of a sort, unlike Lucy Steele's switch of what she calls her "affections" from Edward to Robert Ferrars in *Sense and Sensibility*. In some cases "amour-goût", rather than being interest transmuted into love, is love constrained by interest. Rather than first seeking out the richest potential spouse available and then persuade oneself of being in love, one may simply eliminate from consideration all partners below a certain level of wealth and then let one's affections settle where they may.

The latter scenario could also, however, be due to the imperfection constraint: it is harder to persuade oneself that one is marrying for love if the chosen person is the very richest among all who might be available. The consistency constraint may also operate in such cases. Whereas Lucy Steele had no difficulty transferring her affections to one brother when the other was disinherited, someone who genuinely falls in love with a person who was originally singled out on grounds of interest may not be able to fall out of love when interest dictates that course of action. In a person who has internalized the idea that love is not love which alters when it alteration finds, self-respect may win out over self-interest. As in the case of fairness, there is an asymmetry between adopting an attitude when and because it serves one's interest and giving it up when and because it no longer does. In an entirely speculative vein, we may imagine that in *King Lear* both Burgundy and France love Cordelia because of her prospects, but that only the former cares so little about his self-image that he is able to shed the emotion when it no longer coincides with his interest.

Passion into interest. We may also observe the converse transmutation, that of passion into interest. This is especially likely to happen in contexts, such as market competition, where adherence to the canons of instrumental rationality is vital. To act against a business rival out of envy or a desire for revenge is rarely profitable. Both motives go together with disregard for consequences, in the form of a willingness to suffer a net loss, as long as the other also suffers or suffers more. Price cutting may be an effective form of revenge against a competitor, but the firm practicing it may drive itself out of business as well. One member of a cartel may be so envious of the large profits made by another member that he triggers the breakup of the cartel by lowering prices. When passion, as in these cases, induces actions against one's interest there may be a tendency to persuade oneself that they do in fact promote it. Initiation of a price war may be reconceptualized as an instrumentally rational response to the behavior of the other firm. The head of the firm that breaks out of the cartel may persuade himself that the short-term superprofits will offset the long-term losses.

Although I have not seen any evidence that envy and revenge, dressed up as profit-maximization, actually serve as motives in business transactions, I do not think the idea can be excluded. When two motives – interest and

passion – point in different directions, the tension may be reduced by one of the motives aligning itself with the other. The examples cited above of passion aligning itself with interest, although not compelling, seem quite plausible. The presently cited examples of interest aligning itself with passion are clearly more conjectural.

Passion into passion. Much of world literature turns on the transmutation of one passion into another. I shall illustrate the issue by the transmutation of love into hate in Stendhal's *Le Rouge et le Noir.* In this novel, the daughter of a wealthy and high-placed aristocrat, Mathilde de la Mole, falls in love with Julien Sorel, the son of a carpenter and her father's secretary. Initially, she tells herself that it "shows a high heart and a daring spirit to love a man so far beneath [her] in social position" (II.xi). She views her love for him as something heroic and out of the ordinary. "If, while still poor, Julien had been noble, my loving him would be nothing more than a vulgar act of folly, a commonplace misalliance; I wouldn't want such a thing; it would have nothing of what characterizes a grand passion – the enormous obstacles to be overcome, the dark uncertainty of the outcome" (II.xii). Her love for a social inferior, at this stage, is a source of pride rather than shame because she frames his situation as one of *extraordinary* inferiority rather than of extraordinary *inferiority.*

Giving in to this contrived or artificially heightened passion, Mathilde writes a letter to Julien expressing her love, and then regrets it bitterly: "She was writing *first* (what a terrible word) to a man in the lowest rank of society. This circumstance, were it to be discovered, ensured her everlasting disgrace. Which of the women who came to see her mother would dare to take her side? What tactful phrase could be found for them to say to soften the blow inflicted by the frightful contempt of the drawing-rooms?" (II.xiv) Yet even this acute feeling of social shame is swept away when Julien, playing hard to get, refuses to respond unambiguously to her overture. She summons him to find a ladder and climb into her room. He obeys; she receives him; but almost immediately has second thoughts:

> I've given myself a master, so Mademoiselle de la Mole was saying to herself, plunged into the most doleful grief. He may be the soul of honour, well and good: but if I provoke his vanity to extremes, he will revenge himself by making the nature of our relations known. Mathilde had never had a lover before, and at that moment in life which gives even the hardest hearts some soft illusions, she was tormented by the most bitter reflections
>
> He has tremendous power over me, since he rules by terror and can inflict a frightful punishment on me if I try him too far. This idea was enough of itself to incline Mathilde to insult him, for courage was the prime quality of her character. Nothing could stir her in any way or cure her of an underlying feeling of boredom (*ennui*) constantly springing to life again, except the idea that she was putting her whole existence at hazard. (II.xvii)

At this stage, the two lovers are "unconsciously animated by the keenest hatred towards each other" (II.xvii). When after various vicissitudes they reunite, Mathilde once again breaks with Julien:

> Remorseful virtue and resentful pride made her, that morning, equally unhappy. She was in some sort shattered by the dreadful idea of having given certain rights over herself to a mere

> humble priest, who was a peasant's son. It's almost, she said to herself, as if I had to reproach myself with a partiality for one of the footmen. With proud and daring characters, there is only one step from anger against oneself to fury with other people, transports of rage afford one intense delight in such circumstances [...] Mathilde [...] found it an exquisite satisfaction for her pride in punishing in this way both herself and him for the adoration she had felt a few days before. (II.xx.)

In Mathilde's meta-emotions there may indeed be an element of remorse or guilt, but there is above all social shame and the anger or "resentful pride" triggered by "having a master". While her uncompromising character will accept nothing short of absolute love, which for her is defined as total subjugation to Julien, it also makes her rebel against the thought of actually having him as her master. When Julien in his misery makes the fatal mistake of pleading for himself, he offers her only a pretext for punishing him. At the very beginning of their relationship, she told herself that "at the first sign of weakness I see in him, I'll give him up" (II.xi). What now transpires is the opposite: she justifies her desire to give him up by the weakness he displays when she makes it known to him. She delights in the feeling of righteous anger and the catharsis of revenge.

Mechanisms. The phenomena I have discussed are not all equally robust. The transmutation of interest and passion into reason are well-documented phenomena, the transmutation of interest and passion into one another less so. Transmutations of passion clearly occur, although not all of us have experienced them as dramatically as Mathilde de la Mole and Julien Sorel. Yet whatever doubt one might cast on the reality and importance of this or that class of cases, the general idea is, I believe, hard to deny. By contrast, the mechanisms behind the phenomena are not at all well understood. *Exactly how* does it happen that people fool themselves into thinking that they do what they do for other motivations than those which really animate them? We need a theory or a mechanism-generating framework that can explain the role of the emotions as (i) motive forces of transmutation, (ii) inputs to transmutation and (iii) outputs of transmutation. The account would also have to be capable of explaining transmutations of interest into reason. *But there is no such theory or framework.* Neither of the two dominant theories of motivated attitudinal change, Festinger's theory of cognitive dissonance and Freud's theory of defense mechanisms, addresses the issues I have identified.

A satisfactory theory of transmutation would have to incorporate two simple ideas. On the one hand, people have a strong desire to promote their material interests. On the other hand, they have a strong desire to maintain a positive self-image. For most people, the self-image includes a belief that they are not motivated only or even mainly by material interest. For some, it may include a belief that they are motivated only or mainly by interest. Sometimes, these two desires can be satisfied simultaneously without transmutation. When the oppressed fight for their liberation, interest and self-image are almost inseparable. For the aggressively amoral entrepreneur, too, there is no conflict between the two values. Most people, however, often find themselves in situations where they suggest opposite course of action. Sometimes, they may be able to accommodate both through a process of transmutation. The extent to which they succeed depends on the extent to which they are limited by the consistency and imperfection constraints. Also, some people

may be intrinsically better than others at telling self-serving stories to themselves.

Misrepresentation

The self-deceptive practices discussed above may be contrasted with the less paradoxical phenomena of deception. While a person may be fully aware that he is really animated by one motivation, that very same motivation may lead him to present a different motivation to others. (I shall ignore the psychic strains and tensions that may be caused by this dual attitude, and assume that the person successfully and costlessly maintains a double accounting system of his motivations.) Also, the person may be guided by two different aims: the desire to achieve a certain aim and the desire to be perceived in a certain way by others. Even if the first aim can be achieved without misrepresenting his motivation, the second might, if important enough, induce him to falsify it, at the expense of the first aim.

The target of misrepresentation may either be an interlocutor or a third-party audience. Consider for instance a husband and a wife bargaining over custody as well as over the property settlement. The husband may try to misrepresent himself to his wife as having a strong interest in custody, to get a favorable property settlement. If this is unfeasible, because it is common knowledge between him and his wife that he does not really want to take responsibility for the child, he may address a third party instead. In front of a judge, he may successfully misrepresent himself as having a strong interest in custody, because any statement by his wife to the contrary will be discounted as motivated by *her* self-interest (or spite). Because it is often in people's interest to denounce others as motivated by interest, claims to that effect, even when true, may not be credible to third parties. In fact, the claim might backfire – C might think that when it is in A's interest to claim that B is arguing out of interest, A is probably not telling the truth, an implication being that B probably is.

Or consider a legislator arguing *to* other legislators *before* a national audience of voters. He may dress up his interest in impartial language in order to persuade his fellow legislators that his proposal is well-founded. To the extent that legislators are indeed (i) willing to listen to "the mild voice of reason"[7] and (ii) unable to distinguish insincere from sincere appeals to reason, this stratagem may succeed. Even if neither (i) nor (ii) obtains, however, his interest in getting reelected may induce the legislator to couch his proposal in impartial terms, with a wink that is visible to his co-legislators but not to the public. He may adopt what Richard Posner calls "Aesopian language" (Posner, 1982, p.273) because he fears that voters might punish representatives who explicitly present legislation as pieces of self-interested bargaining. I return to that issue below.

Interest as reason. In many public debates many speakers are mainly motivated by some form of interest. In all public debates, all speakers represent themselves as being motivated reason. In theory, perhaps, the second of these statements is consistent with the first statement being *universally* true. Each and every instance of public reason-giving might represent self-interest in disguise. But there is a problem: *Why bother to argue* if people are universally motivated by self-interest? In response, one might imagine a society in which some members falsely believe that some (unknown) members are genuinely

7. The phrase is by Madison (*The Federalist* No.42).

open to argument and offended by naked appeals to interest. Because the former do not know who the latter are, they may decide to use impartial language on all occasions, on the principle "can't hurt, might help" (see below, however, for ways in which the constraints on impartial language *may* hurt the speaker). Even the third category – someone who is neither open to argument nor believes that anyone else is – might decide to use impartial language, reasoning as follows: "If I express my interest directly without impartial garbs, those who believe in the existence of members open to argument might punish me, because if they don't those who are (falsely believed to be) open to argument might punish them for not punishing me."

This story, although implausible, illustrates an important point. We can change the story and assume that it is common knowledge that some members of society, perhaps a small fraction, are genuinely open to argument, but that one cannot tell who they are. This knowledge might trigger the cascade of simulations just described. When A interacts with B before a public consisting of C,D..., A has two reasons for misrepresenting his interest as an impartial appeal to reason. First, B might, for all A knows, be one of the genuinely impartial members of society. If A expresses his interest directly, he will both lose the chance of persuading B and run the risk of being punished by B. Second, even if B does not belong to that subset, A knows that B will know that one of C,D... belongs to it. A knows that B knows that if B fails to punish A, one of C,D... may punish B, which gives B has an incentive to punish A for expressing purely self-interested concerns.[8] Hence, the known presence of *some* genuinely reasonable people in the population may induce others to mimic their behavior.

Above I gave some examples of self-serving appeals to reason to illustrate the mechanism of transmutation. Several of these might equally plausibly be understood as instances of misrepresentation. When large parties argue for majority voting in single member districts on the ground that this system favors the impartial values of governability and stability, their real motive is often to favor party interest, as their rivals will not fail to point out. When a parent argues for custody by citing the interest of the child, he or she may engage in self-conscious misrepresentation of a purely private interest, as the other parent will not fail to point out. When, in a recent Norwegian wage bargaining case, ambulance drivers cited bank functionaries as a "natural" reference group for wage comparisons, nobody believed they believed what they said.

A more general class of cases arises in legislatures. When an interest group obtains legislation that is passed for no other reason than the benefits created for the group, this fact is rarely stated. Instead, as Posner said, Aesopian language is used. A law requiring the licensing of shoe salesmen, for instance, might be justified on public-health grounds, by the need to limit the spread of athletes' foot.[9] Jonathan Macey argues that "special interest legislation is [..] often drafted with a public-regarding gloss [...] because this gloss raises the costs to the public and to rival groups of discovering the true effect of the

8. There is no reason to go to a higher-order argument, and assume that B might expect a purely self-interested C to punish him because if he doesn't one of D, E.. will punish C for not punishing B. As the transaction between A and B takes place in full view of all other members of society, B knows that he will punished by someone for not punishing A.
9. Posner (1982), p.286. He cites this as an extreme example.

legislation. This, in turn, minimizes the major cost to the legislator of supporting narrow interest group legislation – the loss of support from groups that are harmed by the legislation – and thus reduces the cost to special interest groups of persuading the legislature to vote for the special interest legislation." (Macey, 1986, p.251).

The cost to the interest group of this public-regarding gloss is that courts will give statutes their public rather than their private meaning whenever the two diverge. Since legislative bargains will therefore be only imperfectly enforceable, fewer will be struck.

Another possible cost to the interest groups arises from the imperfection constraint. To fool the public they may have to accept legislation that is somewhat suboptimal from their point of view, provided it is superior to the status quo. Hence the wedge between the private and the public meanings would be present from the very beginning, and presumably create even more opportunities for courts to favor the latter over the former. Before proceeding to give some examples, I shall make some general comments on the place of the consistency and imperfection constraints in conscious misrepresentation.

The need to satisfy these constraints follows from the more general need to prove one's sincerity, i.e. to show that one is not choosing impartial arguments a la carte in a purely opportunistic manner. Thus a speaker may also try to convey sincerity by playing the Devil's Advocate rather than being an out-and-out proponent of his favored policy. "Everything that furnishes an argument against the thesis being defended by the speaker, including objections to his own hypothesis, becomes an indication of sincerity and straightforwardness and increases the hearer's confidence." (Perelman and Olbrechts-Tyteca, 1969, p.457). Pointing out the weaknesses in one's own position has two opposite effects. Although it makes the audience more disposed to examine the argument seriously, it could also make it aware of weaknesses that it might otherwise not have noticed. Rational self-interest would make the speaker lean over backwards, but not too far, in pointing out the weaknesses in his own argument.

Trying to prove one's sincerity by respecting the imperfection constraint creates a similar dilemma. The need for this constraint arises because perfect coincidences arouse suspicion. "Plebiscites and elections yielding results too favorable to the propositions or candidates of the government side have rarely been regarded as a sincere expression of the voters' opinion." (Perelman and Olbrechts-Tyteca, 1969, p.473). Stupid dictators get themselves reelected by a majority of 95%; smart dictators content themselves with 70%. Yet even that majority may be too high to fully deflect suspicion: only loss of the election would be truly convincing. More generally, arguing for a position that deviates from one's first preference has two opposed effects. While making the proposal more credible and thus more likely to be adopted, it also serves one's interest less well if it is adopted. If proposals can be varied continuously, there may exist a policy arguing for which maximizes expected utility – unless the audience is so skeptical that only counterinterested arguments will convince them. If proposals are naturally lumpy or discrete, the closest alternative to one's preferred policy may be so distant that the status quo is preferable. In that case, one is better off not proposing anything: one is damned if one does and damned if one doesn't.

Let me offer four local-justice examples in which policy proposals based on interest succeed because and to the extent that they embody that interest imperfectly rather than

perfectly. The first two cases illustrate negative discrimination, the last two positive discrimination. I do not claim that positive discrimination is always based on group interest, or even that it is invariably in the two cases I shall discuss. Typically, I believe, coalitions behind affirmative action include some who believe that such policies are intrinsically fair and some who support them on the basis of group interest. For the first subset, the misrepresentations that I shall discuss belong to the "reason into reason" subdivision discussed below. For the second subset, they belong here.

(i) Consider first restrictions on the right to vote. In many societies, property has been used as a criterion for the suffrage. One may, to be sure offer impartial arguments for this principle. At the Federal Convention, Madison argued that the stringent property qualifications for the Senate, rather than protecting the privileged against the people, were a device for protecting the people against itself (Farrand, 1966, vol.I, p.421, p.430). But there is something inherently suspicious about such arguments. They coincide too well with the self-interest of the rich. It may then be useful to turn to literacy, as an impartial criterion that is *highly but imperfectly* correlated with property. At various stages in American history literacy has also served as a legitimizing proxy for other unspeakable goals, such as the desire to keep blacks or Catholics out of politics (Creppell, 1989). These cases are, as noted below, more accurately seen as misrepresentation of passion as reason.

(ii) American immigration policy has also used literacy as a proxy for criteria that could not be stated publicly.[10] Proposals to screen immigrants by testing them for literacy in their native language were usually justified as a way of selecting on the basis of individual merit, a widely accepted impartial procedure. The real motivation of the advocates of literacy was, however, usually prejudice or group interest. Patrician nativists wanted to exclude the usually illiterate immigrants from Central and South-Eastern Europe (passion or prejudice misrepresented as reason). Labor feared that an influx of unskilled workers might drive wages down (interest misrepresented as reason).

(iii) Turning now to positive discrimination, recent decades have seen a conflict between the goal of favoring ethnic minorities in college admission and the principle of color-blind admission imposed by courts and state legislatures.[11] Although some colleges have tried to get around this problem by various forms of subterfuge, the current trend is to admit minority students as part of a preferential admission of students from culturally or economically disadvantaged backgrounds. On the one hand, this policy will admit some students from (say) poor Irish families that would not have made it into college under the earlier system, and deny admission to some middle- or upper-class minority students that would have been admitted under a race-based system of affirmative action. On the other hand, it will admit more minority applicants than under a purely merit-based system. By diluting the goal, it becomes more feasible to implement it.

(iv) Consider, finally, the use of race as a criterion of allocating kidneys for transplantation.[12] Three facts conspire to make American blacks badly placed in this allocative pro-

10. The following draws on G. Mackie (1995).
11. The following draws on P. Conley (1995).
12. The following simplified exposition draws on M. Dennis (1995).

cess. First, they are overrepresented as patients. Second, they are underrepresented as donors. Third, because of their unusual antigen patterns they are less likely to benefit from a kidney taken from a white person. To the extent that kidneys are allocated on the basis of antigen matching (an impartial criterion of efficiency), blacks do badly. To compensate for this form of medical bad luck, the United Network for Organ Sharing allows transplantation centers to use time on the waiting list (an impartial criterion of fairness) as an additional principle. This criterion is viewed as acceptable because it does not uniquely favor blacks, but also enhances the prospects of other individuals or groups with unusual antigen patterns.

Interest as passion. The strategic misrepresentation of interest as passion can occur when the agent wants others to believe that he is blind or deaf to consequences. A threat is non-credible when it would not be in the interest of the speaker to carry it out. Because it is known that passion can induce people to act against their interest, the speaker has an interest in presenting himself as moved by passion rather than by interest. Others may then be deterred by the threat because they believe he would actually carry it out. Conversely, a speaker may find it in his interest to appear as moved by passion in order to make others believe that he will not be deterred by a threat that would be effective when directed against a person moved by rational self-interest. "A seeming madman, therefore, may be a superior strategist, because his threats are more readily believed. Could Colonel Ghadafi and Ayatollah Khomeini have understood this principle better than the cool, rational leaders of Western nations trying to deal with them?" (Dixit and Nalebuff, 1991). Hitler, apparently, was a master of this form of deception.[13] Richard Nixon deliberately cultivated an appearance of erratic behavior, in order to persuade the Soviets that he could not be counted on to react rationally to a first strike (Isaacson, 1992, p.181-182). He even, paradoxically, boasted of the fact.[14]

The consistency and imperfection constraints impose sharp limits on this strategy, however. To convey a believable impression that one is irrational, occasional grandstanding is not enough. One has to engage in seemingly emotional behavior on numerous occasions, important as well as unimportant, to create the impression that the irrationality is a character trait rather than a mask. Also, one has to show in practice that one is willing to suffer considerable losses because of one's emotional disposition. Someone who gets out of control when and only when he would gain *on that occasion* by having others believe him irrational will find it difficult to build a reputation for being dangerously emotional. In particular, he will not be credible if he backs off in encounters with other emotional

13. "Hitler never said anything, even when he appeared to have lost his temper, without calculating the effect both on those present and on those to whom they would recount it." (A. Bullock, 1991, p.571).

14. *Ibid.*, p.294. It is probably a mere accident that this episode occurred just one year before the publication in *The New Yorker* of a drawing that shows a disgruntled-looking man selling pencils on the street with a whip in his hand and a sign around his neck saying "Irrational". R. Frank (1988) reproduces this drawing and adds that "the sign round the man's neck is not the only, or even a very good signal that he is not fully rational. On the contrary, that he seems to have realized the sign might serve his purposes can only detract from its ability to do so". Isaacson, by contrast, characterizes Nixon's boasting of his irrationality as "disarming and alarming". My hunch is that Isaacson is right: as a rational man would understand that boasting of his irrationality is self-defeating, doing so is actually self-confirming. The argument could of course, be taken one step further and so on ad infinitum, leaving the issue essentially indeterminate, which may have been enough for Nixon's purpose.

persons. In the language of game-theoretic biology, such behavior will reveal him to be a "Bluffer" rather than a "Hawk" (Maynard-Smith, 1982). Again, there is a trade-off: complete credibility is not desirable if the losses you have to incur in order to be seen as emotional are greater than the gains you can expect from being seen as emotional.[15]

Passion as reason. Given the impulsive nature of many emotional reactions, it would seem that an agent thus motivated would be too caught up in the situation to have the detachment of mind required to misrepresent his motivation. Yet there are at least two classes of cases in which this seems to happen quite frequently. In the first place, the agent may explain his *past* emotional behavior as really motivated by reason (or some other motivation: see below). In the second place, he may be subject to a *standing passion* or prejudice that does not interfere with the capacity for strategic misrepresentation. Here, I illustrate the second case by the misrepresentation of passion as reason. Below, I illustrate the first case by indicating how agents may misrepresent their past emotional behavior as motivated *by another emotion* than the one which actually was at work.

The cases I shall consider turn on the misrepresentation of *prejudice.* Above I mentioned how the literacy test for voting or immigration has served as an impartial pretext for discrimination on ethnic and religious grounds. Another striking example is the policy adopted by Yale College in the 1920's to limit the admission of Jews. Following a recent scandal at Harvard, Yale did not want to use explicit quotas. Instead Yale adopted a policy of geographical diversity, ostensibly as a goal in its own right, but in reality to reduce the number of Jewish students. "Though many individual Jews (concentrated in the northeast region from which Yale received most of its applications) would be affected by this principle, it was not an innately anti-Jewish principle. A geographical policy applied without regard for religion that would help an individual Milwaukee Jew or Duluth Catholic as much as it would hurt a New York atheist or Hoboken Protestant could not appropriately be termed religiously biased."[16] The policy, in other words, satisfied the imperfection constraint, the impartial criterion of geographical diversity serving as a diluted and therefore more acceptable equivalent of ethnicity.

In the United States of the 1990's, racial prejudice can be presented much more simply as an impartially grounded objection to affirmative action. The situation is somewhat similar to the issue of prejudice towards blacks or AIDS patients discussed above. In that case, prejudice allied itself with an outcome-oriented impartial theory emphasizing public health against a rights-based impartial theory emphasizing protection of civil liberties. Here, prejudice allies itself with a conception of impartiality as color-blindness against a veil-of-ignorance conception of impartiality which requires compensation for disadvantages due to color. I am not saying that all advocates of color-blindness are prejudiced, nor can I point to clear cases in which this advocacy is a mere pretext. Yet I believe that those who hold the color-blind view on genuinely impartial grounds would agree that they often find themselves with strange bedfellows; in fact, if they didn't I would suspect

15. Thus after the invasion of Cambodia in 1970, "Nixon's 'madman' strategy was backfiring on him: he was coming across as unhinged in the eyes of his own nation. As a result, the Cambodian invasion would turn into the greatest victory for Hanoi since it lost the 1968 Tet offensive" (Isaacson, 1992, p.270).

16. D. Oren (1985, p.198); see also Conley, (1995).

their impartiality. Glenn Loury, a black conservative critic of affirmative action, certainly agrees when he writes that "selling these positions within the black community is made infinitely more difficult when my black critics are able to say: 'But your argument plays into the hands of those who are looking for an excuse to abandon the black poor'; and I am unable to contradict them credibly." (Loury, 1995, p.21). I mentioned earlier that reason and interest may form coalitions *for* affirmative action; similarly, reason and passion may form coalitions *against* it.

The recent constitutions in Central and Eastern Europe contain impartially worded clauses whose origin is unambiguously found in ethnic prejudice. All the constitutions in the region include clauses that ban (negative) discrimination on grounds of race, nationality, ethnicity, sex, religion and many similar grounds. Only three constitutions – those of Bulgaria, Romania and Slovakia – also contain explicit bans on reverse or positive discrimination, i.e. affirmative action. These are also the countries in the region with the largest minority populations[17] and the strongest history of ethnic conflict. In the Romanian document, the ban only covers reverse discrimination on ethnic grounds. Bulgaria and Slovakia did at least try to satisfy the imperfection constraint by extending the ban to *all* the criteria that are enumerated in the bans on negative discrimination. Yet in these countries, too, the clauses are due to the prejudices of an ethnic majority in the constituent assembly against various minorities. The biases against ethnic minorities would have been even stronger had not delegates from the Council of Europe intervened in the constitution-making processes. The first draft of the Romanian constitution, for instance, contained an impartially worded ban on ethnically based parties that was directly aimed at the large Hungarian minority.

Passion as passion. Certain emotions are more objectionable than others. Because different emotions may give rise to similar behaviors, people may have an incentive to substitute a more acceptable emotion for the one that actually moved them to act. I shall illustrate this idea by some examples from politics in ancient Greece, where the unavowability of envy and *hybris* induced commoners as well as kings to present actions thus motivated in a different light.

In "On envy and hate", Plutarch writes that "men deny that they envy [...]; and if you show that they do, they allege any number of excuses and say they are angry with the fellow or hate him, cloaking and concealing their envy with whatever other name occurs to them for their passion". In Classical Athens, this tendency was revealed by the practice of denouncing others, who claimed to act for the sake of revenge, of being really motivated by envy. In David Cohen's summary, the orator Lysias "argues that his opponent will falsely claim that he brings the prosecution out of enmity so as to get revenge, but in fact it is only out of envy because the speaker is a better citizen. [..] The desire for revenge apparently would be seen by the judges as a legitimate reason for prosecuting, so the speaker must deny that this is the case. Meanspirited envy, on the other hand, reflects badly upon the accuser's character and indicates that the suit is unreliable." (Cohen, 1995, p.82-83).

In ancient Greece, *hybris* – humiliating another merely for one's own pleasure – was a

17. The percentages are: Albania, 2%; Bulgaria, 15%; the Czech Republic, 5.5% (not counting Moravians); Hungary, 8.6%; Poland, 2%; Romania, 10.5%; Slovakia, 14.4%. Source: J. Bugajski (1994).

punishable offense and, moreover, a strongly disapproved form of behavior. Against accusations of *hybris* therefore, it was expedient to represent one's behavior as motivated by a more acceptable urge. In the *Politics*, Aristotle tells a story about a tyrant, Archelaus, who was killed (among other reasons) because one of his boyfriends decided that their association had been based on "*hybris* not on erotic desire".[18] He then offers the advice to tyrants who want to stay in power, that in their acquaintances with youth, they should appear to be acting from desire rather than from *hybris*.[19] In other contexts, those accused of *hybris* represent their behavior as motivated by revenge. Although there may be truth in their allegations of having been wronged, the *hybris* consist in the revenge being disproportionate to the offense.[20]

Passion as interest. In some cases, the more acceptable misrepresentation of an emotion may be interest rather than another emotion. An example that comes to mind is the misrepresentation of fear as prudence. In many societies fear in the face of danger is seen as dishonorable. One can imagine two reasons why this might be so. First, if we are dealing with genuine fear rather than the "aseptic" fear that merely involves a belief that something may happen and a desire for it not to happen, fear is dishonorable because it shows a lack of self-control. Second, either variety of fear might be subject to social disapproval because it testifies to an excessive concern with mere survival and a corresponding lack of concern with honor. In the latter case, the agent has nothing to gain from misrepresenting his fear as a form of prudential behavior. On the contrary, in "honor societies" such behavior is interpreted as a sign of cowardice. In the former case, however, an agent might be motivated to present an image of himself as someone who flees danger out of rational prudence rather than out of panic.

Prejudice, too, may be dressed up as interest. Thus when prejudiced members of a racial or ethnic majority argue against affirmative action favoring minorities, they may take the low road of interest rather than the high road of reason. In societies that are both permeated by interest groups and dominated by egalitarian ideologies, there certainly attaches less opprobrium to interest than to prejudice. Moreover, by defending their views in terms of interest rather than reason, groups can avoid the costs that arise from the consistency and imperfection constraints. In ethnic conflicts, this mechanism may coexist with that described in the previous paragraph. On the one hand, an ethnic group may misrepresent behavior really caused by fear as grounded in rational prudence. On the other hand, it may claim that the conflict is generated by a conflict of interest over scarce territorial resources rather than by hatred or prejudice.

Reason as interest. A person who is genuinely motivated by impartial concerns may find it expedient to argue in terms of self-interest, for one of two reasons. In the first place, he may try to persuade an interlocutor to adopt his proposal by arguing that it is in the interest of both. In the second place, he may appeal to interest if the society in question penalizes appeals to reason. I shall discuss these cases in turn.

Several writers have argued that a just social order is that which would be chosen by rational, self-interested individuals behind a hypothetical veil of ignorance. This basic idea

18. *Politics* 1311b.
19. *Ibid.* 1315a; see also Cohen (1995, p.145) and Fisher (1992,p.30-31).
20. Fisher (1992, p.509) (summarizing his earlier analyses, notably in Ch.XI).

can be spelled out in different ways, to support utilitarian theories no less than the theories of John Rawls and Ronald Dworkin. In none of these versions does it amount to a claim that justice can be deduced from rationality alone. The veil of ignorance, in these theories, is itself derived from prior normative conceptions of what features of individuals count as "morally arbitrary". If we consider actual rather than hypothetical veils, however, mere self-interest may be sufficient to generate consensus on basic constitutional issues, if the relevant outcomes lies so far into the future that nobody can tell for sure how they and their descendants will be affected. This is the structure of a veil-of-ignorance argument that was used repeatedly at the Federal Convention, most strikingly in an intervention by George Mason:

> We ought to attend to the rights of every class of people. He had often wondered at the indifference of the superior classes of society to this dictate of humanity & policy, considering that however affluent their circumstances, or elevated their situations, might be, the course of a few years, not only might but certainly would distribute their posteriority through the lowest classes of Society. Every selfish motive therefore, every family attachment, ought to recommend such as system of policy as would provide no less carefully for the rights and happiness of the lowest than of the highest orders of Citizens. (Farrand, 1966, p.49).

Reason as reason.[21] It may happen, in a given society, that *specific* impartial arguments become suspect. In that case, impartially minded speakers may have an incentive to substitute another impartial argument for the one that has fallen into disrespect. There are two possibilities, depending on whether the speakers do or do not believe in the substitute argument. I shall discuss these two cases in turn.

To illustrate the first case, I can report from a recent public meeting in New York City where I heard a black law professor discuss affirmative action policies with considerable anguish. Although he was clearly in favor of such policies, on grounds of fairness, he also reported that in the current political climate, explicit advocacy of affirmative action generated so much "toxicity" that he would make this argument only if all else failed. For the time being he found it more expedient, he said, to make a substitute impartial argument in terms of support for the economically and culturally disadvantaged more generally. From what he said I inferred that he also believed in the justice of this policy. While he obviously thought that members of racial and ethnic minorities had *stronger* claims than disadvantaged members of the white majority, he admitted that the latter, too, had some claims on society.

I shall illustrate the second case with a debate from October-November 1789 in the French Assemblée Constituante. At that early stage of the revolution, the delegates had not yet adopted the ruthlessly consequentialist attitude that came to dominate them in later stages, memorably enshrined in the Comité du Salut Public. It was far from being generally accepted that established rights could be overridden for the sake of the common good;

21. The misrepresentation of *reason as passion*, while perhaps conceivable, will in general be too inconsistent with Kantian and Habermasian norms to be a coherent behavior.

in fact, any such argument was sure to be met with disapproval. Utilitarian or efficiency-oriented framers, therefore, were constrained to frame their arguments in terms of rights. This is what happened in the debates in the Assemblée Constituante over the confiscation of Church property.

In their attempts to justify the confiscation of the Church goods, both the opportunistic Mirabeau and the hypocritical Talleyrand argued that these goods in reality belonged to the nation, instead of simply saying that the acute financial crisis made this measure necessary. The argument, unbelievably bad it was, went as follows. If the Church had not, on the basis of its income and property, provided religious services and assistance to the poor, the State would have had to do so. Therefore, the State is the real owner of that property and no rights would be violated by turning it over to the State. The best reply to this specious argument came from Clermont-Tonnerre, based on a deep and modern understanding of the rights of corporate actors. But we can also follow Camus and proceed by analogy. A father has the obligation to provide a dowry for his daughter. Assume that a friend or a relative is willing to provide it instead, thereby discharging the father of his obligation. Should we imply that he thereby becomes the owner of the dowry offered to his daughter? Although I find it hard to believe that anyone in the assembly believed that those who made the rights-based arguments for confiscation believed in what they were saying, the proposal was adopted.

Motivations and constraints. The reasons for disavowing a certain motivation may be intrinsic or extrinsic. On the one hand, if the motive was known it might induce disgust and contempt in others that in turn would induce feelings of shame in oneself. The disguise of fear as prudence, or of envy as anger, illustrate this case. On the other hand, public knowledge of one's motivation might be counterproductive in terms of that motivation itself. The disguise of interest as reason or passion, of passion as reason, and of reason as interest, illustrate this case.

In all cases, the misrepresentation is subject to the consistency and imperfection constraints. The consistency constraint, in addition to the costs it may impose by forcing us to act against our interest, also imposes the purely mental costs involved in keeping track of one's lies: "What a tangled web we weave when first we practice to deceive". Montaigne wrote that "My wit is not supple enough to dodge a sudden question and to escape down some sideroad, nor to pretend that something is true. My memory is not good enough to remember that pretense nor reliable enough to maintain it: so I act the brave out of weakness. I therefore entrust myself to simplicity, always saying what I think". Or as he also says, "Even if I did not follow the right road for its rightness, I would still follow it because I have found from experience that, at the end of the day, it is usually the happiest one and the most useful."

References

Babcock, L. and Olson, C. 1992, "The causes of impasses in bargaining". *Industrial Relations* 31:348-60.

Babcock, L. Wang, X. and Loewenstein, G. 1996, "Choosing the wrong pond". *Quarterly Journal of Economics* 111:1-20.

Babcock, L. et al. 1995 "Biased judgements of fairness in bargaining". *American Economic Review*, 85:1337-43.

Bugajski, J. 1994 *Ethnic Politics in Eastern Europe.* London: Sharp.

Bullock, A. 1991 *Hitler and Stalin.* New York: Vintage Books.

Cohen, D. 1995 *Law, Violence and Community in Classical Athens.* Cambridge University Press.

Conley, P. 1995, The allocation of collage admissions. In Elster J. 1995a.

Creppell, I. 1989 "Democracy and literacy: The role of culture in political life". *Archives Européennes de Sociologie* 30:22-47.

Dennis, M. 1995, Scarce medical resources: Hemodialysis and kindly transplantation. In Elster J. 1995a.

Dixit, A. and Nalebuff, B. 1991 "Making threats credible", in R. Zeckhauser (ed.), *Strategy and Choice.* Cambridge, Mass.: MIT Press.

Elster, J. 1989a *Solomonic Judgements.* Cambridge University Press.

- 1989b *The Cement of Society.* Cambridge University Press.
- 1992 *Local Justice.* New York, Russel Sage.
- 1993 "Rebuilding the boat in the open sea: Constitution-making in Eastern Europe". *Public Administration* 71, 169-217;
- 1994 "Argumenter et négocier dans deux assemblées". *Revue Française de Science Politique* 44, 187-256;
- 1995a *Local Justice in America,* New York, Russel Sage.
- 1995b "Transition, constitution-making and separation in Czechoslovakia". *Archives Européennes de Sociologie* 36, 105-34.
- 1995c "The impact of constitutions on economic performance", in *Proceedings from the Annual Bank Conference on Economic Development,* Washington; The World Bank, pp.209-226;
- 1995d "Forces and mechanisms in the constitution-making process". *Duke Law Review* 45 1995, 364-96; - "Strategic uses of argument", in K.Arrow et al. (eds), *Barriers to the Negotiated Resolution of Conflict.* New York: Norton, pp.236-57.

Farrand, M. 1966 ed., Records of the Federal Convention, New Haven: Yale University Press.

Fisher, N.R. 1992 *Hybris.* Warminster: Aris and Phillips.

Isaacson, W. 1992 *Kissinger,* New York: Simon and Schuster.

Loewenstein, G. et al. 1993 «Self-serving assessments of fairness and pretrial bargaining», *Journal of Legal Studies* 22:135-59.

Loury, G. 1995 *One by One From the Inside Out.* New York: The Free Press.

Macey, J. 1986 "Promoting public-regarding legislation throug statutory interpretation: An interest-group model", *Columbia Law Review* 86:223-68.

Mackie, G. 1995, "U.S. immigration policy and local justice". In Elster J. 1995a.

Maynard-Smith, J. 1982 *Evolution and the Theory of Games.* Cambridge University Press.

Montaigne. 1991 *The Complete Essays.* translated by M.A. Screech, London: Allen Lane.

Oren, D. 1985 *Joining the Club: A History of Jews and Yale.* New Haven: Yale University Press.

Perelman, C. and Olbrechts-Tyteca, L. 1969 *The New Rhetoric.* University of Notre Dame Press.

Posner R. 1982, "Economics, politics, and the reading of statues of constitution", 1982 *University of Chicago Law Review* 49:263-91.

Rabin, M. 1995 «Moral preferences, moral constraints, and self-serving biases». Unpublished manuscript.

Sniderman, P. Brody, R. and Tetlock, P.E. 1991 *Political Psychology.* Cambridge University Press.

Swenson, P. 1988 *Fair Shares,* Cornell University Press.

Veyne, P. 1976 *Le pain et le cirque,* Paris Seuil.

Zajac, E. 1995 *The Political Economy of Fairness,* Cambridge, Mass.:MIT Press.