

COGNITIVISM IN POLITICAL PHILOSOPHY

S. L. Hurley

In Morality and Well-Being: Essays in honour of James Griffin, edited by Roger Crisp and Brad Hooker, Oxford University Press, 2000.

“Η γνωσιοκρατία στην πολιτική φιλοσοφία” (“Cognitivism in Political Philosophy”, Greek translation by Nicos Stavropoulos), Isopoliteia 1998.

1. Introduction. This essay provides a bird's eye view of a cognitivist approach to political philosophy, by illustrating its application to three topics: justice, democracy, and punishment. The arguments are merely sketched rather than given in detail, to give an overall picture of the approach.¹

Cognitivism in political philosophy is not a doctrine or thesis, but a category. Cognitivist accounts are cast primarily in terms of truth and knowledge rather than choice or preference. There are quite different ways of falling under this broad characterization, as we'll see. But some unity can be given to a broadly cognitive approach by two key ideas.

A constraint on political cognitivism is that it give some basis for responding to certain worries. These worries prompt political liberals such as Rawls to deny that the search for truth can provide a shared basis for a conception of justice in a pluralistic democratic society.² On this view, a pluralistic democratic state that avoids the authoritarian use of state power should be neutral about the conflicting and incommensurable conceptions of the good held by its citizens. Now Rawls is not a sceptic about ethical truth. But he still sees the absence of commitment to ethical ideals, even liberal ideals like autonomy, as essential to liberalism as a political doctrine. There is no practicable answer for political purposes to the question of the true good, since public agreement on this cannot be obtained in a nonauthoritarian, pluralistic, democratic society.³

How can these concerns be met within a cognitivist approach? This is where two key ideas come into play. One is that we can often know that certain biasing influences tend to undermine knowledge, even when we make no politically controversial assumptions about the truth. Moreover, we don't need to know what precise relationship between truth and belief makes for knowledge in order to know that certain factors tend to defeat knowledge. For example, even if we don't think that the notion of beliefs tracking the truth is all epistemology needs, we may recognize that knowledge can be defeated by influences on belief that could not possibly vary counterfactually with the truth, such as desires to believe certain things. Given lack of public agreement and the difficulty of identifying positive expertise, political cognitivism considers how we can nevertheless seek knowledge of the answers to political questions, by at least avoiding the biasing influences that make knowledge impossible. Note that authoritarian power is itself a biasing influence; those who hold it tend to surround themselves with people who tell them what they want to hear.

A second key idea is more positive: we must develop and use effectively certain capacities of citizens. The cognitive capacities of individuals are valuable both in their own right and to the social search for knowledge of what should be done. We can recognize that general cultivation of the cognitive, deliberative and ethical capacities of citizens are

necessary for certain democratic procedures to avoid bias, even if we don't know the truth about the questions those procedures address.

James Griffin has made a persuasive case for the need to identify beliefs of special reliability in practical and ethical reasoning. I share with him the view that such beliefs include the core value beliefs that provide conditions of intelligibility. But I am less sceptical than he about the possibility that procedures and institutions can be designed, even in the absence of ex ante knowledge of the further truths at issue, which tend to lead to reliable beliefs. He writes:

...to say that we should interest ourselves only in judgements formed in the absence of conditions likely to corrupt judgment begs the important questions. If we knew which conditions did that, and also knew we were avoiding them, we should indeed be able to isolate a class of especially reliable judgments. (Griffin 199 , p.)

But conditions likely to corrupt judgement can be hard to recognize; self-interest is a master of disguises.

However, these difficulties are greatest in the personal sphere, when cognitive labor cannot easily be divided. When faced by a difficult personal decision, we cannot appeal to institutions that delegate the decision to the least biased decision makers, or take a crash course in the virtues we need to make the decision well. But in the public sphere, a political distribution of cognitive labor has a distinctive contribution to make. Political institutions and procedures can be designed and adjusted to avoid overall bias and foster capacities in the long term, while minimizing controversial political presuppositions. These two principles give a cognitive twist to the concerns that motivate political neutrality without actually being committed to neutrality. So they are well-suited to play a key role in perfectionist versions of liberalism. They at once make cognitivism politically palatable in nonauthoritarian pluralist societies and gives politics the potential to be of cognitive value.

These two key ideas will be illustrated by brief sketches of how they can be developed in the three areas mentioned. Both ideas will be illustrated in relation to democracy⁴, bias-avoidance in relation to justice, and the fostering of capacities in relation to punishment. Note that in section 2 we discuss the just distribution of goods, while in section 3 we discuss the democratic distribution of cognitive labor.

2. Justice. Theorizing about distributive justice often seems to be guided implicitly by the assumption that a basic aim of egalitarianism is to neutralize the effects on distribution of brute luck, understood as factors for which we are not responsible. (Here we'll use 'luck' as short for 'brute luck'; to neutralize luck in this sense is to track responsibility.) This assumption has been made explicit and clarified in important recent work by Cohen (1989, 1992) and by Roemer (1985, 1986, 1987, 1993). Luck in this context is usually understood to include luck in the kind of person you are, such as genetic luck, which supposedly gives rise to what Rawls (1971) calls "morally arbitrary" natural and social differences between people. The aim to neutralize the influence of such factors on distribution can be seen as having an implicit methodological role in Rawls' theory, in providing the normative significance of choices made in the Original Position. This is the case despite Rawls' official aim to avoid making issues of desert and responsibility prior to issues of justice. In Cohen's and Roemer's

work, the role of luck-neutralization or responsibility-tracking is more explicit and less methodological, that of a basic substantive aim of justice.

It may be natural to assume that rejection of the aim to neutralize luck has anti-egalitarian consequences. For example, while arguing against Rawlsian egalitarianism, Nozick implies that you do not have to deserve to **be** everything that you are in order to deserve the results of what you **do** (1974, p. 225). By contrast, the view taken here is that rejection of the luck-neutralizing aim in its familiar methodological or substantive roles can strengthen rather than weaken egalitarianism. The negative reasons for this suggestion must be passed over here (see Hurley 1993, 1995.) Some hints: our concepts of responsibility and luck are not determinate enough to do the work needed by some luck-neutralizing accounts of justice. When someone is not responsible for what he has got, there may simply be no answer to what he is responsible for instead. Moreover, the aim to neutralize the **effects** of luck is tied to the suspect principle that in order to be responsible for something you have to be responsible for its **causes**. This principle makes responsibility impossible, because no one can be responsible for the causes of everything, all the way back. But the supposition that no one is responsible for anything provides no support for the claim that everyone does deserve the same thing.

There are also positive reasons for departing from a luck-neutralizing view: perhaps surprisingly, a stronger case for egalitarianism can be made by assuming a different fundamental aim, which displaces responsibility from centre stage in theories of distributive justice in favor of knowledge. Egalitarianism should give a central role not to the aim to neutralize **luck**, but instead to the aim to neutralize **bias**. We can admit that there is an important general connection between justice and responsibility. But this connection can play a background role, the role of setting certain parameters for incentive-seeking. By contrast, in contemporary theories of justice this connection has tended to play central roles, whether methodological or substantive, whether explicitly or implicitly (Cohen 1989).

Biases are influences that distort the relationship of our beliefs about what should be done to any truths there may be about what should be done. It is antecedently unlikely that biassed beliefs will constitute knowledge--unlikely, that is, independently of any view about the truth or falsity of the beliefs in question. Biases are cognitive distortions: they distort the relationship of belief to truth in a way that prevents belief from attaining the status of knowledge. For example, a personal desire to believe that something is true is a biasing influence on belief. If you believe something because you want to believe it, then even if your belief just happens to be true, it isn't knowledge. Someone whose beliefs on a topic just happen to be true in this way doesn't have beliefs that are reliably true under relevant counterfactual suppositions. That his own beliefs on this topic are not reliable may mean that they are not a reliable general source of input to deliberation about the truth. Many but not all desires distort the formation of beliefs and thus are biasing influences, and so does certain information that bears on how such desires can be met.

Bias is issue-relative. The same factor can be a biasing influence with regard to certain issues, but not others; love of your spouse, for example, can make for more accurate insights about his character. When a factor is biasing for one issue but not another and both are in the public domain, weighing and balancing is needed. Perhaps a degree of patriotism motivates citizens to the cognitive efforts democracy requires, for example, but too much may distort beliefs about the relationship of your own country to others.

We should distinguish two strands of thought about justice, which are often run together. In Rawls' theory, for example, there is a Kantian, luck-neutralizing strand. This plays a central methodological role within his theory. It is implicitly expressed by his concern to eliminate arbitrary natural and social contingencies from the Original Position, and his remark that the Original Position can be thought of as the point of view from which noumenal selves see the world (1971, p. 255). But there is also a distinguishable, arguably cognitivist strand, which can be understood in terms of a concern to neutralize bias.

The latter, cognitivist strand can take us further toward an egalitarian account of distributive justice when detached from the former, in part because it then does not inherit the problems about responsibility. Aiming at knowledge provides a reason to adopt a perspective of ignorance in thinking about how goods should be distributed: because ignorance of ourselves would rule out many biasing influences, such as those deriving from self-interest. Egalitarianism can be supported from such a perspective in a way that does not depend on the luck-neutralizing aim.⁵

Consider a generalization of the Rawlsian framework for thinking about distributive justice. We describe some normatively significant, fictional point of view from which principles of distribution are to be derived, under certain constraints. From such a constrained perspective, we consider the question: "What should be done about distribution?" We do not assume that the answer is necessarily to be motivated by self-interest operating within the given constraints. The general idea is that, if we specify the right, normatively significant constraints on such a point of view, principles of distribution derived from it will be fair, will be principles of justice.⁶ But this could be either because, as in Rawls, the procedure determines what count as fair principles, or because we think it is likely that principles resulting from such a procedure will yield justice--or at least more likely that they will do so than that principles arrived at in other ways will. We can call such a point of view a perspective of justice. This is a generalization of Rawls' idea of the Original Position for two reasons. First, because it does not assume an exercise of constrained self-interest must occupy this perspective. Second, because it does allow that the perspective may be a "device" in aid of the discovery of just principles rather than a way of determining what count as just principles (compare Scanlon 1982, 122).

We can now further distinguish two ways of setting up the perspective of justice. One, often appealed to by utilitarians, involves assuming you have an equal chance of being anyone in society. That assumption, among others, acts as a constraint on the reasoning to principles of justice. This is the equal chance characterization of the perspective of justice. It assumes there are known probabilities or risks. The other way of setting up the perspective of justice is Rawls' way: to assume instead a different constraint, namely, that you are radically ignorant of who and what you are. This is the ignorance characterization of the perspective of justice. It does not assume that risks are known, but merely assumes ignorance or uncertainty.

These assumptions are distinct so long as you do not adopt the principle that lack of information concerning different possible outcomes justifies assigning them equal probability. Some orthodox decision theorists favor this principle, but Rawls explicitly resists it. And experimental subjects tend to reject the inference from ignorance to equal probability implicitly, by reacting differently to situations involving uncertainty as opposed

to known equal probabilities (an example follows). That is, most people are far more averse to situations of uncertainty, where risks are unknown, than they are to running known risks. And there is now a flourishing school of nonexpected utility theorists who do not dismiss this tendency as irrational.⁷

An example may help to see the difference between risk aversion and uncertainty aversion. As an empirical matter, most people are risk averse when they know the probabilities involved. This means that if you offer people the chance to play a game in which they have an objective 50% chance of winning \$100, most will pay less than the game's actuarial value of \$50, though gamblers will pay more. Consider a game where someone wins \$100 if she guesses correctly the color of a ball drawn from an urn containing 50 red and 50 black balls. It is worth something to play this game; how much? A typical answer, expressing a degree of risk aversion, might be \$30. If you now change the game so that there are 100 balls, red or black, but their proportions are unknown, you are now dealing with uncertainty rather than risk. And typical offers drop dramatically, to around \$5 (Raiffa 1961). This difference reflects the difference between risk aversion and uncertainty aversion. They are logically independent; a gambler who likes to take known risks could still be averse to acting without knowledge of risks, under uncertainty.

The distinction between the equal chance and ignorance characterizations of the perspective of justice can be related to several other distinctions: to the distinction between risk aversion and uncertainty aversion, and to the distinction between the luck-neutralizing aim and the bias-neutralizing aim. On the one hand we have a set of ideas involving risk and luck: the aim to neutralize luck, the assumption that I might have an equal chance of being anyone, and the idea of risk aversion. On the other hand we have a set of cognitive ideas involving ignorance and knowledge: the aim to neutralize bias, the assumption that I am ignorant of who I am, and the idea of aversion to uncertainty. Consider how various permutations of these ideas might work in derivations of principles of distribution from a perspective of justice.

First, how does the ignorance vs. equal chance issue interact with the risk aversion vs. uncertainty aversion issue? While as a matter of fact most people are risk averse, Rawls does not want to assume, in the course of reasoning for his principles of justice, that the parties to his Original Position are risk averse. This is because he holds that different attitudes to risk should be respected as part of people's differing conceptions of the good, which are veiled in the Original Position. And if risks are unknown in the Original Position, risk aversion cannot not get a grip anyway; only aversion to uncertainty is relevant.

By contrast, there are very general conceptual reasons to assume that, other things equal, intentional agents prefer information to lack of information, so are at least weakly averse to uncertainty. Consider the alternative: to accept uncertainty, other things equal, would be to decide how to act without taking account of available information and reasons. That would be hard to reconcile with the minimal cognitive rationality and capacities we require of intentional agents. So we can justify on conceptual grounds adding an assumption of weak uncertainty aversion to the ignorance characterization, even if we cannot justify adding an assumption of risk aversion to the equal chance characterization. We'll return to this matter.

Second, how does the ignorance vs. equal chance issue interact with the bias-neutralizing vs. luck-neutralizing issue? Should the hypothetical position from which principles of justice are chosen be understood in terms of ignorance of who we are or rather of equal chances of being anyone? Rawls of course favors the ignorance assumption. But what do the aims to neutralize luck or bias have to say on this question?

The aim to neutralize bias supports the idea of deciding about principles of justice in a position of ignorance of your own advantages and disadvantages and your probabilities of gain or loss from various options. Knowledge of these matters might well be biasing, whatever the truth about justice. For example, how much truth is there in the view that offering large incentives to the most talented is necessary to get them to produce in a way that benefits everyone (see Cohen 1992)? We don't need to know the answer to this hard question in order to recognize that my beliefs about this may be biased by knowledge of whether I am talented or not, and my corresponding desires to believe one thing or another. You can be biased because you're talented and want very much to believe the talented need large incentives to produce in a way that benefits the less talented, while I am also biased because I'm not talented and want very much to believe the talented do not need large incentives to benefit the less talented. Anyone is more likely to reach the right answer if she can abstract from these biasing influences.

The aim to neutralize bias also argues against the idea of deciding about justice on the basis of calculations of your chances of gain. But this holds even if these calculations derive from an assumption that everyone has equal chances. Biases can distort beliefs even if they apply equally to everyone.

To elaborate this point: The aim here is to remove information that would allow potentially biasing desires to affect deliberation about what should be done. In this sense, desires relating to your own chances of gain can be just as biasing as desires relating to your own certainty of gain. But this point is not affected by everyone's chances of gain being the same. The kind of bias in question is cognitive: a distortion of the ideal relationship between truth and belief, introduced by certain desires. It is only derivatively bias in the sense of partiality, or differential orientation to or concern for one person as opposed to another. Bias in the cognitive sense may operate when your beliefs are influenced by desires that operate via calculations of your chances of gain, even if everyone's chances of gain are the same so that impartiality obtains. Cognitive bias is distinct from partiality: the equal chance characterization can admit bias, expressed in calculations of your chances of gain, even though it excludes partiality. For these reasons the bias-neutralizing aim does not lend itself to conceptions of the perspective of justice as yielding expressions of constrained self-interest, to the extent the constraints in question still admit cognitive distortion. In this respect the position taken here has an affinity with Scanlon's criticisms of the reading of contractualism in terms of rational self-interest operating behind a veil of ignorance.⁸

There are further reasons, related to the bias-neutralizing aim, that also favor an ignorance characterization of the perspective of justice over an equal chance characterization. A fundamental question here is this. Should we model our attitudes to the distribution of resources across persons, all of whom are or will be actual, on our attitudes to the distribution of resources across possible states of affairs, not all of which will be actual? (Compare Scanlon 1982, p. 127.) This is what we do when we base an answer to the question of which actual people should get what on the supposition that I have an equal

chance of being anyone. The analogy provides a decision theoretic device for reducing interpersonal problems to individual decision problems given risk. But the formal parallel between a life being mine and a possibility being actual is disturbing in the context of thought about justice. For example, we want to say something like: "So what if the chances of it being my handicap are small? It's still somebody's real handicap; the relevance of some actual person's actual handicap to considerations of justice just doesn't depend on the chances that it might have been mine." To the extent attitudes to uncertainty do not reduce to attitudes to risk, the ignorance characterization fares better than the equal chance characterization in this respect. Imposing ignorance allows the deliberative decision maker to focus on the human reality of the handicap or other relevant circumstance, by avoiding the bias that goes with knowing that she herself is safe--or probably safe.

How does the aim to neutralize luck bear on the ignorance vs. equal chance issue? Let's apply the luck-neutralizing aim to luck in the kind of person you are, or constitutive luck. Consider the idea of a natural lottery of constitutions, or sets of essential properties, the outcome of which is morally arbitrary. Suppose that in such a lottery you have an equal chance of being anyone. This assumption can be regarded as giving a very literal reading to the aim to neutralize constitutive luck. That is, making a decision in the ex ante position where there are equal chances of various possible constitutions might seem a way to neutralize such luck.

But this literal reading is in danger of incoherence: who or what is the 'I' who has an equal chance of various different sets of essential properties? We should not want to be committed to making sense of such constitutionless selves. So the luck-neutralizing aim does not support the equal chance characterization, at least not in this way.

Does the luck-neutralizing aim instead support the ignorance assumption? Whether it does or not, we don't need it. The bias-neutralizing aim can support the ignorance assumption without help from the luck-neutralizing aim. Notice that the reasons why the bias-neutralizing aim supports the ignorance assumption are quite independent of the luck-neutralizing aim and worries about the moral arbitrariness of natural advantages or disadvantages, and they carry no implications about constitutionless selves. Here the negative reasons we passed over for avoiding a luck-neutralizing account are also relevant: since we don't need the luck-neutralizing aim to support the ignorance over the equal chance characterization, we might as well avoid the problems about responsibility that aim generates.

The net result is that the aim to neutralize bias supports the idea of deciding on principles of justice from a hypothetical position of ignorance rather than one of equal chance. To get this far with Rawls we don't need the aim to neutralize luck, with all its attendant problems.

However, we admitted that there is an important general connection between justice and responsibility. We can now get a glimpse of how this can be given a background, parameter-setting role, rather than a central role. How equally resources are distributed by a maximin principle depends critically on what assumptions we make about incentive-seeking (Cohen 1992). If the most productive seek little or no incentives for being highly productive, maximin will yield a very egalitarian outcome. But if the most productive seek large

incentives, maximin may, under further assumptions, allow very substantial inequalities for the sake of raising the level of the worst off a rather small amount.

What assumptions about incentive-seeking should we make? Natural and sociological forces may set a range of possible incentive-seeking behaviors that are realistically feasible in modern conditions. But it is reasonable to assume that within that range, social and ethical norms influence levels of incentive-seeking. In particular, norms concerning how responsible people are for the results of what they do will have an important influence on the levels of incentives that are needed to avoid demoralization and economic apathy.

Both the truth about responsibility and popular intuition about responsibility lie between extremes. It is neither true that people are responsible for almost everything they do nor that they are responsible for almost nothing they do. Incentive-seeking behavior is sometimes and to some extent expressive of a natural and widespread sense of moderate responsibility for what someone does, as a result of his abilities. But this natural sense of responsibility does not license unrestricted incentive-seeking. To merit praise and reward for your activities is not to merit any reward you can extract, however exorbitant. The natural sense of responsibility is vague, but carries with it an unavoidable sense of proportion; it is not the view that you deserve anything you can get away with. But equally it is not the erroneous view that no one is responsible for anything they do as a result of their abilities because no one is responsible all the way back for the causes of those abilities.

It is a good thing for everyone that this latter, erroneous view is not widely held, for several reasons. By making responsibility seem impossible this view would undercut the critically important wellbeing that derives from our sense of responsibility (Strawson 1986, p. 87). Moreover--and this is the important point for egalitarians--it would undermine this wellbeing unselectively, across all socio-economic classes. A view that undercuts the sense of responsibility is a blunt instrument. Its use for egalitarian purposes is handicapped by its universally negative consequences for the wellbeing that flows from our sense of responsibility.

These vague truths about responsibility set the parameters of incentive-seeking behavior within the range of feasibility: an assumption of moderate incentive-seeking should be made, corresponding to the vague, moderate popular sense of responsibility. Individuals do need some significant portion of their product as incentive to produce, given this norm of responsibility. But they do not require that almost the whole of their extra product be returned to them. We should not look for more determinacy here than the vague truths about responsibility admit of. Nor do we need to, so long as we keep responsibility in this background, parameter-setting role in relation to a general assumption of moderate incentive-seeking. We do not need to decide in particular cases what goods someone does deserve, when he is not responsible for what he actually has. By avoiding such decisions we avoid occasions for bias to outstrip truth.

Notice that despite any methodological role luck-neutralizing may be given in justifying the Original Position and the assumption of ignorance, these general background issues about responsibility and incentive remain. Until responsibility has played its parameter-setting role in relation to incentive-seeking, the implications of a maximin principle for distributive justice are indeterminate. The real work done by responsibility is

here in the background. But this background work does not require that the luck-neutralizing aim motivate the perspective of justice. The bias-neutralizing aim does a better job in that central methodological role.

Where do we go from here? Rawls wishes to derive a maximin principle of distributive justice from the Original Position, which requires inequalities in primary goods to benefit the worst off groups in society. So, for example, inequalities resulting from incentives to produce given to the most talented would only be permitted if they benefitted the worst off. Now one way to derive a maximin principle from a perspective of justice would be to ignore the distinction between ignorance and equal chance and to assume risk aversion. In the absence of risk aversion, an equal chance assumption will not generate egalitarian principles. Moreover, we'd have to assume extreme risk aversion to get a maximin principle for resources from an equal chance assumption. Otherwise, reasoning from the equal chance assumption, I might be tempted to trade risk-avoidance off against my chances of great wealth by allowing further inequalities.

But this is not Rawls' way. He does distinguish ignorance and equal chance. Risk aversion would get no grip under the ignorance characterization, since risks are unknown (and risk aversion is logically independent of uncertainty aversion). In any case, Rawls gives good reasons not to assume risk aversion, as we've seen. People's attitudes to risk vary, as well as their tastes and conceptions of the good. Their different attitudes to risk should be respected; no particular attitude to risk should be enshrined in the Original Position. But how can the rabbit of maximin be pulled out of the hat of ignorance under these conditions?

We cannot discuss Rawls' attempt to do so here. But not all have been persuaded by it. We will try a different way, using an assumption of aversion to uncertainty instead of an assumption of risk aversion. We only need weak, not extreme, aversion to uncertainty in order to derive egalitarian principles of distributive justice from a position of ignorance--though how weak determines how egalitarian.

Isn't this just as bad as assuming risk aversion? No: the same objections do not apply. This is not because uncertainty aversion is more marked and widespread than risk aversion, as an empirical matter, though that is true. As we've seen, there is a very general conceptual reason to assume that, other things equal, intentional agents prefer information to its absence. Our conception of minimally rational intentional agents requires them to be weakly averse to uncertainty, even if in order to avoid bias we put such agents in a hypothetical position of ignorance to choose principles of justice. This combination of uncertainty and aversion to uncertainty yields egalitarian results. And it does so in a way that does not give a motivating role to the aim to neutralize luck or to worries about the moral arbitrariness of natural advantages or disadvantages.

A little uncertainty aversion goes a long way in generating egalitarian results. If you start with equal chance, you need extreme risk aversion to get a degree of egalitarianism that you can get instead from ignorance plus weak risk aversion. If ignorance of your identity and traits is not translatable into known chances of gain or loss, it is not possible to trade off avoidance of uncertainty against your chances of gain. But then such trade-offs are ruled out without assuming extreme uncertainty aversion. Recall that it was for a closely related reason that the bias-neutralizing aim favored the ignorance assumption over the equal chance assumption. Part of the point of imposing ignorance, according to the bias-

neutralizing aim, is that it gives no scope to calculations of your chances of gain. But by the same token, neither does it give scope to trade-offs of chances of gain against aversion to either risk or uncertainty. So ignorance combines with even weak uncertainty aversion to preclude the influence of desires that relate to your own chances of gain.

However, the ignorance assumption does allow certain desires to operate. First, an aversion to uncertainty may be regarded as equivalent to a desire for information about the way things actually are or will be. Unlike desires relating to your own chances of gain in particular, this desire cannot in general be regarded as biasing in the sense of cognitive distortion. Second, we need to depend on some general assumptions about basic goods, in something like the way that Rawls depends on the notion of primary goods. Basic goods may include the realization of capacities and the sense of responsibility. Consider then what we can call the 'Pareto preference': an impersonal and general desire that people be better off in terms of these basic goods rather than worse off, other things equal. This simply makes operational our necessary assumption of some basic goods, so does not introduce bias: if it did introduce bias, then those goods would be the wrong ones to assume. So uncertainty aversion and the Pareto preference are compatible with the bias-neutralizing aim in a way that concern for your own chances of gain is not. Moreover, uncertainty aversion can be assumed even though we do not assume self-interested motivation.

Consider then a perspective of justice characterized by the ignorance assumption, by weak uncertainty aversion, and by the Pareto preference relative to the assumed basic goods. In this situation, a maximin principle can be seen as the closest you can reasonably come to avoiding uncertainty over fundamental goods. We can reach this conclusion in several steps, by exploiting the connection between uncertainty and the distribution of goods.⁹

First: Even if you have no information about your position, distributing a known fixed quantity of goods equally avoids uncertainty over those goods. You can easily calculate what everyone actually gets; there is no uncertainty about people's absolute or relative positions.

Second: If the quantity of goods for distribution is not known or fixed, it's no longer possible to avoid uncertainty about absolute positions. But distributing goods equally at least avoids uncertainty about relative positions.

Third: But now remember that uncertainty avoidance is weak, not extreme. Consider setting against it not any known chances of gain for yourself, but rather increases in basic goods for some unknown members of society, so long as the level of the worst off is kept as high as it can be. Moving from equality to maximin admits uncertainty about relative as well as absolute positions, in order to satisfy the Pareto preference. But this is compatible with an assumption of only weak uncertainty aversion. You are not here trading off uncertainty avoidance for an increase in your own chances of gain. Rather, you are trading it for increases in some unknown persons' level of basic goods, increases that leave no one worse off. You judge that their improvement outweighs the additional, relative, uncertainty. This is not because you calculate that you have some chance of being them. Their improvement engages the Pareto preference directly. (Again, compare Scanlon 1982.)

There is a further question as to exactly where the compromise between uncertainty avoidance and Pareto improvements should be struck. Should weak Pareto improvements be permitted, which benefit some and make no one worse off, including of course the worst off? If so, uncertainty about who will benefit is admitted. This will result if we keep the assumption of uncertainty aversion very weak, so that uncertainty should be avoided only other things equal. Or should only strong Pareto improvements be permitted, which make everyone better off--again, including the worst off? If so, at least we know everyone will benefit. A stronger assumption of uncertainty aversion will move us in this direction, but it does not need to be extreme. The first view would admit more relative uncertainty for the sake of Pareto improvement than the second view would. But neither will permit inequalities that involve anyone's being worse off than they would be if we stopped with the unknown total amount distributed equally at step two.

Notice the way the ignorance constraint and the aversion to uncertainty work together. It is only in the context set by the ignorance constraint that aversion to uncertainty drives us toward equality. But aversion to uncertainty itself supports the ignorance constraint, just because the constraint rules out biasing influences, cognitive distortions. So ignorance can serve knowledge. Given the ignorance constraint, we would still (other things equal) like to avoid as much uncertainty about actual distributions of goods as possible, and by keeping everyone equal at an unknown absolute level, we do. If we allow some people to rise above this level but do not know who they are, we sacrifice some knowledge about relative levels. In particular, we do not know which people will be above this level or even the chances that particular people will be above this level. But it is precisely this knowledge about who would benefit that the ignorance constraint has denied us, in the interests of avoiding cognitive distortion. The trade-off we make at step three between uncertainty avoidance and Pareto improvement is in harmony with the underlying aim of bias-neutralization. Ignorance of the identity of those persons who will be above the floor of stage two is precisely the kind of ignorance that avoids cognitive distortions, such as those of self-interest, or envy, that would be introduced if we considered the chances that we would be among those persons.

On the cognitivist view, the normative significance of the perspective of justice characterized by ignorance rather than equal chance is found in the aim to avoid the biasing influences that inevitably go with information, even probabilistic information, about who you are and what you are like, your talents and handicaps. And weak aversion to uncertainty can be justified as part of the minimal cognitive rationality required for intentional agency. By giving a cognitive slant to the perspective of justice, and dissociating it from the luck-neutralizing aim and related ideas, we can reclaim egalitarian results. In this way we can take a lead as egalitarians from Rawls even if we do not share his Kantian, luck-neutralizing sympathies. And this way of developing political cognitivism has not at any point implied authoritarian power or threatened the pluralist and democratic character of liberal society.

3. Democracy. Let's now move on to a somewhat briefer sketch of a cognitivist approach to democracy. Consider two prevalent assumptions about liberal democracy. First, people should vote on the basis of their preferences rather than their beliefs about what should be done. Knowledge about what should be done is not the aim of democracy. The aim is rather to satisfy preferences, which of course requires a way of resolving conflicts between them. This noncognitivist assumption dominates social choice theory, which is

concerned with how to aggregate individual preference orderings. Second, the justification of political decisions and procedures should be neutral with respect to the conflicting conceptions of the good held by different citizens, and should not derive from the aim to learn the truth about how such conflicts should be resolved.

These two assumptions contrast with a conception of liberal democracy as deliberative and developmental, such as that found in J.S. Mill (1958), which can be regarded as a form of liberal perfectionism. On one version of such a view, democratic procedures and institutions should meet two conditions.

First, they should embody a distribution of cognitive labor (or, in this section, distribution for short). This assigns issues to decision-makers or procedures, such as to individuals in their private capacities, to referenda, to the judiciary, to ministers, to the legislature, etc. Moreover, it aims to do so in a way that avoids subjecting the overall decision-making exercise to biassing influences, influences that make it antecedently unlikely that the exercise will yield knowledge. As we've already seen, biassing influences can often be identified even when the truth at issue is not prejudged: consider vested interests, self-deception, wishful thinking, prejudice, propaganda, common inferential error, etc. Even though positive expertise may be difficult to identify without prejudging issues, the distribution can still seek to avoid bias.

Second, democratic procedures and institutions should foster and develop the capacities of citizens for practical knowledge, in particular the capacities for public and private deliberation and for the formation of unbiased belief. The cognitive capacities of individuals should be fostered as well as the cognitive capacities of the socially distributed system of cognitive labor. The quality of public deliberative exercises may depend on the capacities of citizens for autonomous deliberation in the face of conflicting goods, and on the education and opportunities for exercise these capacities need in order to be developed and realized. A cognitive conception of democracy gives a critically important role to the education of citizens and makes considerable demands on them, as voters or in whatever capacity they contribute to the democratic decision-making process. It recognizes the cognitive capacities of individuals as valuable in their own right, as well as in respect of their contribution to socially distributed cognition.

The first of these two principles focusses on the cognitive capacities at the collective level, and the second focusses on cognitive capacities at the individual level. We should avoid over-simply assumptions about the relations between these levels.

For example, the first of these two principles must be interpreted so as to respect some important features of social distributions of cognitive labor, some of which are related to features of distributed information processing more generally. A group can perform certain overall cognitive functions even though there is no representation of that overall function or central plan anywhere in the system. Distributed cognition may be more adaptable than centralized cognition, both to external change and to the cognitive failures of individuals who are components of the system ('graceful degradation'). Perhaps the most important feature for present purposes is that groups can have cognitive properties that differ significantly from those of the individuals in the groups. Certain forms of collective decision-making may accentuate individual cognitive properties both in the good and the bad direction (as in Condorcet's Jury Theorem). But the cognitive capacities of different

groups depend on the social organization of distributed cognition as well as on the cognitive properties of individuals. For example, one way or organizing individuals who display a cognitive failing (such as confirmation bias) may inherit this failing from them, while another may not (Condorcet 1986; Hutchins 1995, ch. 4, 5, and esp. pp. 178, 199, 224, 226-227, 348, etc; Hutchins describes a neural network simulation of some of these features of socially distributed cognition).

Suppose we wish to find a political distribution of cognitive labor that, among other things, avoids biasing influences on the cognitive performance of the group. We cannot assume that the only way to do this is to avoid dependence on any individual components that are subject to biasing influences. Some organizations of individuals may inherit bias from individuals while others may not. However, we do need to study the relationship between biasing influences on individuals, various forms of cognitive organization, and the cognitive properties of the group. So we do need to identify biasing influences on individuals, in order to consider whether cognitive labor should be distributed in a way that avoids depending on biased components at all, or whether they can rather be organized in such a way as to neutralize bias at the level of the system as a whole. Pitting one bias against another within a socially distributed cognitive exercise may sometimes be an effective way to unbiased the overall exercise: consider the adversarial legal system as a possible example. The moral is: we need information about the biasing influences on individuals to design an unbiased distribution of cognitive labor, even though we cannot assume that distributed cognition necessarily inherits bias from individual participants.

Subject to these clarifications, our two principles tell us to find a distribution of cognitive labor that avoids bias overall and that fosters the capacities of citizens. These two principles can, for example, guide institutional respect for freedom of speech. Disagreement, open debate and argument are of cognitive value, both in revealing error and flushing out truth, and in forming the autonomous deliberative capacities of citizens. However, protection so motivated might well not extend to forms of expression that systematically damage rather than foster such capacities, as arguably certain types of pornography may. Reasonable adversarial procedures and assignments of procedural rights should be guided by consideration of the cognitive value of such procedures. Pressure groups and lobbyists should not be given a role inconsistent with that value. The variety of opinions and reasoned disagreement in society provide scope for the operation of a distribution of cognitive labor. Such disagreement should be exploited by a thorough distribution of authority that avoids the general potential for bias inherent in concentrations of power per se, as well as the more particular sources of bias stemming from particular interests.

A cognitive theory of democracy does not claim that all nonauthoritarian distributions that foster the autonomous deliberative capacities of individual citizens will be of overall cognitive value. Or even that they are likely to be. Rather, the claim is the other way around. Distributions that are of overall cognitive value will be nonauthoritarian, so more adaptable and less likely than centralized authoritarian arrangements to lead to disastrously wrong answers. By fostering the autonomous deliberative capacities of individuals, a nonauthoritarian distribution of cognitive labor can protect the conditions in which the search for right answers at all levels can continue. So authoritarian regimes that neglect development can be ruled out on cognitive grounds. But it is a further question what nonauthoritarian distribution is best in a particular context. There are many possible

nonauthoritarian distributions of cognitive labor. Certain assignments of issues to decision-makers and institutions will fit a certain society's characteristics, and others will not.

Democratic procedures should be a means to knowledge of truths about what should be done. However, their relationship to values and reasons for action is not wholly instrumental and evidential. We can see this by considering the way in which democratic procedures contribute to realizing one important value, namely, the value of individual autonomy. Participation in democratic deliberation is not wholly separate from the development and exercise of individual autonomy, but partly constitutive of it. The importance of this point becomes clearer when we hold in focus the intrapersonal as well as interpersonal plurality of goods. We should not conceive of social conflict and deliberation as concerning primarily conflicts between persons, each of whom has a unified, comprehensive doctrine of the good.¹⁰ In modern pluralistic societies, very few persons have any such thing; eclecticism is the norm, even in religion. Rather, pluralism reflects conflicts among values that are widely shared. So the conflicts that need resolution cut within people as much as between them, and the intrapersonal plurality of goods provides the basis for exercises of autonomy. In this sense autonomy is a higher-order value: it presupposes and operates on other values. For example, I may be a member of a socialist trade-union and of a highly competitive and meritocratic orchestra, and so need to form a view about the relationship between the potentially conflicting values these affiliations express. For these reasons, an individual's deliberation about the conflicts she faces and public deliberation about social conflicts are intertwined. Democratic procedures for public deliberation contribute to realizing individuals' capacities and the value of individual autonomy. None of this requires that autonomy be of overriding value. It may itself conflict internally or with other important values.

We can apply some of these considerations by reinterpreting social choice results in terms of belief rather than preference. Social choice theory concerns how the conflicting preference orderings of alternatives by individuals can be reasonably aggregated into a social ordering. It turns out, surprisingly, that no possible process of aggregation can meet certain combinations of conditions on reasonable aggregation. For example, Arrow's impossibility theorem tells us that no process of aggregation will guarantee that social rankings meeting certain conditions are not cyclical. These conditions collectively rule out all possible methods of aggregating individual preference orderings into social orderings (Sen 1970.)

Such results take on a new and significant role in designing a democratic distribution of cognitive labor. The attractiveness of some of these conditions depends largely on the noncognitivist framework usually presupposed in contemporary social choice theory. That is, the conditions are conceived to apply to preference orderings rather than beliefs. So the task of aggregation, the reasonableness of which is at issue, is conceived to be a noncognitive one. By contrast, a cognitive view of democracy as involving the aggregation of beliefs provides a principled epistemic basis for rejecting or restricting some of the conditions. So the corresponding negative results about democracy do not follow, given this cognitive reinterpretation. On this view, what is needed is a method of mapping individual or component beliefs onto collective belief; collective belief is some function of component beliefs. What conditions are appropriate to impose on this function?

Consider, for example, one of the conditions required for Arrow's result, the Independence of Irrelevant Alternatives. This ensures that social preference over pairs of alternatives depends only on individual preferences over the same alternatives. So it forbids social preference between any alternatives x and y to vary counterfactually with anything except the preferences of individuals as between x and y . In effect, it says that the impact of individual preferences on social choice between x and y depends only on the ordering between x and y given by those preferences, not on how strong the preferences are, or on their motivation or genesis, on whether they are well-informed or arbitrary, or on any other information about them.

Why might it be reasonable to deny all such further information an influence on social choice? If you suppose there are no truths at issue, and the task at hand is a strictly noncognitive task of preference aggregation, then it might seem that the use of any further information is ultra vires and question-begging, in implying some objective standard or measure for preferences. Whether it is justified to impose the Independence condition on preference aggregation is controversial, even on a noncognitive view of social choice, because of its role in ruling out interpersonal comparisons of strength of preference. But the point here is that a cognitive conception of democratic aggregation provides independent epistemic reason to reject this condition.

If democracy seeks knowledge of what should be done on the basis of the beliefs of citizens in various roles, then it needs some method of aggregating those beliefs. A method of aggregating beliefs, or function from individual to collective belief, amounts to a method of distributing cognitive labor. The conditions reasonable to impose on any such method will be conditions that avoid subjecting the distributed cognitive process overall to biasing influences without, as far as possible, prejudging the truths at issue. We should not assume that the overall cognitive process will necessarily inherit the biases of individuals or other components, as we have seen. Given certain biasing influences on individuals, some distributions of cognitive labor may inherit bias, while others may not. Nevertheless, the method of distribution needs information about biasing influences, and about the relations between component and system biases, or individual and collective biases, in order to distribute cognitive labor in a way that avoids bias overall. So the method of aggregating beliefs/distributing cognitive labor should be sensitive not just to the content of beliefs about the alternatives at issue, but also to external information about the reliability of beliefs from certain sources about certain kinds of issue in certain circumstances. This kind of information is needed to assign issues to subdomains of decision-makers in a way that avoids bias overall. If democratic procedures are to have cognitive value, institutions should be designed to arrive at beliefs that are, as far as possible for a particular society, unbiased, well-informed and well-considered, not arbitrary. Some persons or group of persons or institution or level or branch of government may be a reliable source of such beliefs for certain circumstances and types of issue, and others for others. A method of aggregating beliefs that places equal reliance on all possible opinions irrespective of their sensitivity to the truth cannot include institutional arrangements that function to distribute cognitive labor in accordance with the distinctions of reliability that fit particular societies and circumstances.

In this light, the Independence condition is a nonstarter. Consider how it would constrain a method for distributing cognitive labor. The cognitive analogue of the Independence condition applies to the beliefs of voters and other social decision-makers

instead of their preferences. Take a set of the beliefs of all relevant decision-makers. These beliefs will have given contents about whether x or y should be done. The cognitive analogue of the Independence condition would require that beliefs with these contents have a fixed effect on the social view as between x and y , regardless of any further information about the influences operating on such beliefs. This rules out evaluation of the counterfactual reliability of beliefs from different decision-makers on different types of issue. Some sources may be reliable on some types of issue but open to biasing influences on other types of issue. A method of distributing cognitive labor needs that very information to assign issues to decision-makers, procedures, institutions, and so on, in a way that avoids biasing influences overall. This is true even though the method of distribution does not prejudice the truths at issue. As we have seen, some distributions of cognitive labor may work better than others, for given individual or component cognitive properties. So a method of distribution should be sensitive to these properties. Information about, among other things, the biasing influences on individuals or other components of the distributed cognitive system is needed to distribute labor effectively. But the Independence condition applied to a method of aggregating belief/distributing cognitive labor would deny us this information. The counterfactual sensitivity involved in evaluating bias violates Independence. So Independence cannot be a reasonable condition to impose on the cognitive task of aggregating beliefs with the aim of achieving knowledge. Given a cognitive conception of democracy, the Independence condition would amount to an unreasonable handicap.¹¹

The basic idea here is that the counterfactual structure given to social choice theory by various social choice conditions can be evaluated by reference to the conflicting demands of counterfactual reliability made by knowledge. One of the best known results in social choice theory has been used to illustrate how social choice conditions can be reinterpreted and evaluated in cognitive terms. But related points can be made for other social choice results and conditions.

4. Punishment. Let's now turn to the topic of punishment. Here is a brief sketch of how three strands of thought might be woven together into a cognitivist approach to punishment. The suggestion is not that this kind of view should stand alone in justifying punishment. We already have good reason to think we need a complex overall theory of punishment. But it may well form an essential part of such a complex theory. The three strands of thought are, first, the broadly Aristotelian idea of practical knowledge; second, the long-term character-forming influences of punishment emphasized by Scandinavian theorists such as Andenaes and Ross; and third, the communitarian idea that certain nonneutral social conditions are needed to sustain and realize essential capacities of citizens. The idea of bias avoidance emphasized earlier will not be used here, but the other key cognitivist idea, that of fostering capacities for knowledge, will be.

The first strand is a broadly Aristotelian idea of practical knowledge. This notion involves an ideal of the rationality of emotional responses. It rejects a presumed dualism or conflict between reason and emotion and allows that emotions themselves may be subject to cognitive standards. How should I feel about something? The way a good person, a person of practical wisdom, would feel. The practical knowledge that enables someone's feelings to be sensitive to the real nature of the situation, rather than being distorted or inappropriate, is acquired through practical experience and training. Practical education is a matter of practice, habituation, being properly brought up in a morally civilized community. Alluding

to recent moves in cognitive science that have allied themselves with Aristotle, we could say that practical education is a matter of having your neural networks trained up by interactions with ethical practices in your environment so that you acquire ethical perceptual capacities and skills (Churchland 1995, pp. 148-150). Among the results should be feelings of aversion to evil and feelings of shame, where these reflect the true nature of the situation and are appropriate.¹²

The second strand is an important correction to a gap in many English-language discussions of punishment. This can be found in the work of Scandinavian theorists such as the criminologist Johannes Andenaes and the jurist Alf Ross. Andenaes distinguishes the long-term character-forming influences of the practice of punishment from the shallower influence of deterrence as it is usually conceived. He also emphasizes that, even if long-term formative influences of punishment are strong and important, certain factors will make it difficult to acquire evidence for this. Ross rejects the standard contrast between retributive and deterrence theories of punishment and stresses that the general preventative effect of punishment depends primarily on the way the shame and infamy attached to punishment influences and forms feelings and attitudes in society at large. (See Andenaes 1974; Ross 1975, e.g., p. 89-90.)

Many philosophers have tended to see the influence of punishment in preventing certain evils along the lines of what we can call shallow deterrence. The word 'deterrence' suggests an influence that holds a great deal about the potential criminal and his society constant: you take the potential criminal's personality structure and desires, as well as prevalent social norms, as exogenously given, and you simply alter the payoffs that attach to certain actions by adding penalties to crimes. These penalties may be thought of as the price of the crime. They may be external sanctions, either legal or social, or internal sanctions. But even the internal versions are often thought of as the pricks of conscience, mere internal analogues of the unpleasant external consequences of crime: as if the internalized penalty structure were wrapped around a core character and system of beliefs and desires that are still held constant.

The idea of shallow deterrence, whether it operates externally or internally, can contrasted with the idea of influences that reach right into the constitution of the self, which alter the very character and personality factors that deterrence takes as exogenously given and by reference to which it calculates its penalties. These would be formative influences on people's characters, values, and systems of beliefs and desires, as they grow up and beyond, in part via the ethical and social norms prevalent in society. But criminal law and the legal practice of punishment might well contribute to such long-term formative influences, and through them have general preventative effects on crime. If so, the theory of punishment should not regard personality, character and the capacities of citizens for practical knowledge, as exogenously determined; such factors are endogenous to the theory of punishment.¹³

Andenaes notes that most people wouldn't commit certain crimes, regardless of the threat of punishment. But he goes on to ask how long this would remain true if criminal sanctions were abolished. Might they may be among the influences that form people's characters, in ways that society cannot afford to dispense with? Such influence would not be impossible to overcome. If the law is simply out of touch with social norms, it will be disrespected, as in Prohibition. Nevertheless, the law is a real influence: Andenaes gives the

example of mandatory prison sentences in Norway for driving under the influence of alcohol, and their influence on the development of social mores there. He also describes the way punishment neutralizes the demoralizing consequences of witnessing crime: frustration at seeing the criminal get away with it with impunity, at seeing the bad example reinforced, the tendency to ask why I should restrain myself when others are not, the chain reaction effects of crime, and so on. Bad examples, which might otherwise be imitated, are made less attractive. Much of the influence of punishment may be unconscious, at the level of absorption of taboos throughout childhood and beyond. This may operate in conjunction with mere deterrence: when someone is deterred by the threat of punishment, there may be a kind of psychological dissonance left, which is resolved by derogation of the action: he didn't really want to do it anyway, it was nasty. In this way punishment can prompt unconscious internalization of norms. Also, people have a tendency not to focus on, to blur over, the negative aspects of something they want to do, including criminal acts. Apprehension and the threat of punishment can have an eye-opening effect: Andenaes gives the example of a shop-lifter who has been deceiving himself about his own activity, rationalizing it to himself, not facing up to it. When caught, he suddenly sees clearly what he has been doing and feels terribly ashamed, and his friends are ashamed for him also, and reinforced in their determination never to get themselves in a similar mess. In this way the legal penalty triggers social sanctions. On balance, Andenaes regards the direct short-term moral educational influence of the criminal law to be fairly weak. But he regards its indirect long-term effects in reinforcing inhibitions and influencing character-formation as important. He regards criminal law as one of the fundamental socializing influences, which operates not just through fear, but influences human thought and behavior in varied and subtle ways that permeate society.

Although punishment may in this way have a general preventative influence on criminal behavior, it will most likely be a very long-term influence, often indirect, with feedback: criminal law and punishment may influence ethical norms and character-formation, while ethical norms also influence criminal law. Such long-term influences are extremely difficult to isolate from other influences, especially since they operate through social norms. Even if we assume for the sake of argument that the formative thesis is true, we should expect difficulties in finding evidence for it that meets rigorous scientific standards, for various reasons cited by Andenaes. Our experience of the effects of the criminal law tends to be of people who were not deterred, namely, of criminals, not of people who were deterred. Not only is the true extent of unreported crime of certain kinds unknown, but much more fundamentally, we don't know how many people have been prevented from committing crimes who would have committed them otherwise. It is difficult to gauge the effects of altering punishments, because it is difficult to establish proper controls. Comparing different geographical areas with different penalties requires us to know whether differences in criminality are due to other social differences; but areas with very similar social conditions and sharply different penal systems may be hard to find.¹⁴ Comparisons across time in one society face similar problems: penal changes are usually gradual, and are accompanied by other social changes, so how can you isolate factors to be clear about what causes the final result? If the long-term influences of punishment on criminal behavior take a generation, or several, to operate, in this time many other compounding influences may operate also.¹⁵

But paucity of evidence for the thesis that punishment does have long-term formative influence in preventing crime does not mean that there is evidence for the

contrary thesis, that it has little such influence. That would be just as hard to come by. We can understand why direct evidence may be hard to get either way here. So we need to ask ourselves what is reasonable to assume as a default position and where the burden of proof should be. This point is obvious, yet its implications are often ignored when the lack of evidence for a disputed proposition is adduced.

The Scandinavian formative view of punishment can be compared to Hampton's (1984) moral education theory of punishment, which focusses on the education of the individual punished rather than on general social influences. It can also be compared with Feinberg's (1970) expressive view of punishment, which focusses on the immediate expressive function of punishment rather than on the long-term consequences of such expression. But there isn't space here to pursue these comparisons.

Finally we come to the third line of thought. Communitarians have emphasized that capacities essential to sustain respect for justice and political legitimacy require certain social and cultural conditions. This in itself seems relatively uncontroversial. This category of capacities should include the autonomous deliberative capacity emphasized in the discussion of democracy, as well as the capacity for practical knowledge with its emotional aspects. What are more controversial are the further communitarian claims, first, that institutions governed by liberal neutrality will be incapable of supporting the social conditions necessary to sustain respect for justice and legitimacy, and second, that what are needed are a politics of the common good and state support for a shared form of life. (Taylor, 1985; Kymlicka 1990, p. 216).

Now it is possible to doubt that liberal neutrality is sufficient to sustain respect for justice and legitimacy without agreeing that what is needed is the fostering and protecting by the state of certain shared ways of life. Liberal perfectionism can be seen as accepting the first of these claims but not the second. Even in the absence of allegiance to neutrality, we may worry that the idea of favoring certain ways of life in the context of contemporary pluralistic societies may lend itself to revival of the white-male-club, or of similar hierarchical or exclusionary forms of life: that the commonality would be accomplished by exclusion of those who are different. However, this is not the only alternative to liberal neutrality. The social conditions for the capacities essential to sustaining justice may depart from neutrality without going all the way to state support for a shared way of life.

Consider the capacity for practical knowledge and for rational and cognitively appropriate emotional responses, including an aversion to evil and a sense of shame where it is appropriate. There are some things, after all, that almost all societies have agreed are evil and shameful, even if they have tolerated them, some things that punishment should not be regarded as simply the price of. Respect for reasonable pluralism should not dilute our recognition of violent crimes and the destruction and suffering they produce as evils. There are some truths we are entitled to prejudge, which we can expect legislatures to respect. The socio-legal practice of punishment may have a long-term influence on the formation of citizens' characters, both as they grow up and beyond into adulthood, so that they see such evils as shameful and so that a basic aversion to the evils of violent crime is an essential part of their personalities. Indeed, we might find communitarian language about the social constitution of the self appropriate in thinking about such influences. There are many differences between the family and political contexts. But punishment in both contexts may be integral to the existence of a properly brought up citizenry, with realized capacities

for practical knowledge. If so, the practice of legal punishment may be among the social conditions for the development of capacities essential to sustaining respect for justice. That is, the formative thesis may provide a way of arguing for the communitarian claim.

The Scandinavian formative thesis and the communitarian claim have just been applied to the capacity for practical knowledge and for cognitively appropriate emotions. But is the cognitive aspect of this application doing any work in a justification of punishment?¹⁶ Couldn't parallel points be dissociated from a cognitive view of the capacities that need cultivating? So long as useful desires and emotions are cultivated, why does it matter if they reflect knowledge or cognitive standards?

This question in effect suggests that we can hold onto the distinction between mere shallow deterrence and long-term formative influences, even if the latter apply to desires and emotions only, and not to ethical perceptions and beliefs. But this is implausible. If we are holding much about an agent's character and world view constant, as in shallow deterrence, we can isolate certain desires and tinker just with them through penalties and incentives. But when our aim is to guide the formation of stable, integrated ethical characters in the long term through social practices, perceptions and beliefs do not fractionate from desires and emotions. It is implausible to suppose we could effectively guide the formation of ethical desires and emotions only, in a way that is neutral with respect to ethical perceptions and beliefs. Ethical character-formation is an ambition that involves unity and harmony between cognitive and conative aspects of character, rather than merely noncognitive conditioning. Perceptions of why certain things are wrong should be integral to and in equilibrium with desires and emotions and social skills. The unified, Aristotelian view of practical knowledge makes the formative thesis far more plausible than it would be for purely noncognitive capacities.

Do these suggestions threaten the value of autonomy appealed to earlier, in the discussion of democracy? The threat is illusory, for several reasons. First, autonomy should not be conceived as the absence of formative causal influences. If we so conceive it we make it impossible, hence make fostering and respecting autonomy irrelevant to politics. Second, as we've just seen, we do not aim to condition or manipulate, which would threaten autonomy, but rather to form cognitive capacities. The autonomous exercise of cognitive capacities is not threatened by causal influences on their formation. Third, autonomy is a higher-order value, which depends on the capacities of people to reflect and deliberate on appropriate first-order values. The value of autonomy doesn't depend on the autonomous person's reaching right answers all or even most of the time, but it does depend on his recognizing some basic first-order values as providing *pro tanto* reasons, and on his having the capacity for higher-order reflection and deliberation. The value of autonomy, that is, presupposes a subject with an ethical constitution. We can cultivate the basic first-order values, in part via the formative causal influence of punishment (both within families and in political society), while also fostering and respecting autonomy. In fact, autonomy presupposes such first-order values; in their absence (as in some kinds of mental deficiency) it is ill-defined. The 'autonomy' of a psychopath may for this reason be a sham. There is still plenty of room within these constraints for respecting autonomy whether or not people reach right answers in exercising it.

But the objection may be reformulated. Suppose someone does recognize the normal range of first-order values, but when in a conflict situation he reflects and deliberates he

nevertheless autonomously decides to do an evil act: say, to commit murder. He supposes the evil is somehow overridden by other considerations. If we take an aim of punishment to be to influence the formations of people's characters so that such exercises of autonomy do not occur, do we threaten the value of autonomy?

Here, we face a conflict within the value of autonomy: the evils in question are evils in part because they involve the complete negation of one person's autonomy (the murdered person's, for example) by another person. We can reasonably judge that exercises of autonomy with such evil consequences should be avoided, in the name of autonomy. In such cases, the end of mutual respect for autonomy may be served by means of formative influences that do amount to restrictions on autonomy. But this cost is worth paying, in terms of the value of autonomy itself, since considerations of potential victims' autonomy may outweigh considerations of potential criminals' autonomy. Implicit here is the view that the state cannot properly wash its hands of violations of autonomy merely because they are not the immediate result of state action or because they involve private rather than public acts.

For these reasons a state role may be needed, via penal institutions, in supporting a shared aversion to evil and a cognitively appropriate sense of shame among citizens. These emotional traits do not relate solely to the evil of injustice; they do not respect the boundary between the right and the good. In the absence of such traits, we may doubt that the sense of justice and allegiance to just institutions will be strong enough to sustain respect for justice (cf. Rawls 1993, p. 142). In this way the communitarian charge that it is sociologically naive to suppose liberal neutrality will sustain respect for justice echoes the Scandinavian insistence that it is naive to ignore the formative effects of state punishment as expressing a common sense of shame and aversion to evil. And these traits may be needed to sustain respect for justice and legitimacy even though a shared way of life or politics of the common good are not.

It may be objected that this distinction fails because talk of shame and recognition of evil is a mere correlate of talk of the good. But while the correlation claim is true, the objection is nevertheless facile. Many ways of life and conceptions of the good overlap in recognizing certain evils. The good that consists in the absence of these widely recognized evils does not define a shared way or life or a distinctive or comprehensive doctrine about the good or the good life. Rather, it provides an important shared presupposition of many such conceptions. Public support for a shared aversion to evil and appropriate sense of shame, fully compatible with reasonable pluralism, may be needed to sustain respect for justice and rights, rather than either a shared sense of justice by itself or a shared conception of the good in the rich sense. We don't need to choose between liberal neutrality and a communitarian politics of the common good; the dichotomy is spurious. A reasonable pluralism may itself require the support of a shared aversion to evil and sense of shame.

Something along these lines may temper a variety of approaches to punishment that in different ways treat the offender too much in isolation from his society and his motivations as too private. We have made great advances in respecting the human rights of individual prisoners, and we should not want to go back on these. But at the same time, we may need to regain sight of the way in which the social context provided by the practice of legal punishment can contribute to making us the persons we are, including the vast majority of us who never venture nontrivially into criminality.

In summary: the practice of punishment may, via long-term formative influences, be among the social conditions of the capacity for practical knowledge, including a shared aversion to basic evils and appropriate sense of shame, which may be needed to sustain respect for justice. There are many ways of distorting this suggestion, and many ways for it to go wrong in practice. But the social consequences of trying to do without the formative role of punishment may be undesirable enough, in the form of widespread violent crime, to warrant focusing attention on how to design penal institutions that harness that formative role to appropriate cognitive standards.

5. Concluding remarks. There is scope for cognitivist approaches to other topics in political philosophy, such as the justification of the adversarial legal system and of freedom of speech. Some of the ways in which political cognitivism can internalize liberal concerns have been indicated, though its affinity is with perfectionist rather than neutralist liberalism.

NOTES

For helpful comments and criticism I am indebted to Brian Barry, John Charvet, G.A. Cohen, Roger Crisp, Raymond Geuss, Brad Hooker, Christine Korsgaard, Christopher Kutz, Luis Martin, Derek Parfit, Joseph Raz, Andrew Reeve, Tim Scanlon, Paul Seabright, Stephen Shute, Quentin Skinner, Larry Temkin, Andrew Williams, Chris Woodard, and members of audiences on various occasions when I have presented this material.

1. In some cases the detailed arguments can be found in Hurley 1989 or 1993.
2. His concern is especially with metaphysical and moral truth. Rawls writes that his conception of justice as fairness "presents itself not as a conception of justice that is true, but one that can serve as a basis of informed and willing political agreement between citizens viewed as free and equal persons". Given the existence of "conflicting and incommensurable conceptions of the good", a fundamental feature of modern culture which social theory must recognize, we secure this agreement by trying, "so far as we can, to avoid disputed philosophical, as well as disputed moral and religious, questions. We do this not because these questions are unimportant or regarded with indifference, but because we think them too important and recognize that there is no way to resolve them politically. ... Philosophy as the search for truth about an independent metaphysical and moral order cannot, I believe, provide a workable and shared basis for a political conception of justice in a democratic society." "...[A] conception of the person in a political view...need not involve...questions of philosophical psychology or a metaphysical doctrine of the nature of the self. No political view that depends on these deep and unresolved matters can serve as a public conception of justice in a constitutional democratic state." In such a state, under modern conditions, conflicting and incommensurable conceptions of the good are bound to exist; this fundamental social fact must be recognized by any viable political conception of justice that avoids reliance on authoritarian use of state power. The absence of commitment to moral ideals, even liberal ideals such as autonomy, he regards as essential to liberalism as a political doctrine; there is no practicable answer for political purposes to the question of the true good, since public agreement cannot be obtained. Rawls 1985, pp. 230-231, 245, 249. See also 1993 on why the idea of moral truth is not a suitable basis of public justification, eg p. 129, and the chapter entitled "The Idea of Public Reason".
3. Of course, people may disagree about the acceptability of Rawls' theory of justice: for example, about the normative significance of the Original Position, or about the significance for justice of those features of persons Rawls considers morally arbitrary. They may also disagree about what constitutes a biasing influence. At some level, almost any substantive theory will be open to disagreement.
4. These two key ideas are importantly related to the views of J.S. Mill in 1958, pp. 27-28.
5. The claim is not that the bias-neutralizing aim identifies a distinctively egalitarian motivation. Rather, it is employed quite generally within a cognitive approach to political theory and its consequences for various areas are considered. The claim is that, when applied to issues of distributive justice in the way described, the bias-neutralizing aim supports egalitarianism better than the luck-neutralizing aim does.
6. This claim is not here defended. It has sufficient currency to merit discussion even in the absence of defense. Issues about its justification are of course also of interest, but are not the topic here.
7. See Bacharach and Hurley (1991), Introduction, sections 2, 3, and passim, for a view of the debate; see also, e.g., McClennan 1990; Gardenfors and Sahlin 1982; Hurley 1989, ch. 4 and pp. 368-382.

8. Compare Scanlon's remarks about the difference between appealing to the idea of rational self-interested choice under special conditions and appealing to the idea of what no one could reasonably reject, given that they are seeking a basis for general agreement. 1982, pp. 122-125.

Scanlon comments on the moves, first, from the idea of impartiality to the ignorance characterization of a perspective of justice, and, second, from the ignorance characterization to the equal chance characterization (p. 121). While Rawls criticizes the second move, familiar from arguments for Utilitarianism, Scanlon focuses on the first move, from impartiality to ignorance (p. 124). He distinguishes a valid version of this first move from a problematic one. If I would have reason to accept a principle no matter which social position I were to occupy, "...then my knowledge that I have reason to do so need not depend on my knowledge of my particular position, tastes, preferences, etc." (P. 121.) This is a valid train of thought, but it does not lead to the different notion of rational self-interested choice under special conditions. Asking what I could agree to in ignorance of my true position is a "corrective device" for "considering more accurately the question of what everyone could reasonably agree to", and does not reduce to the quite different idea of what would be chosen by a single self-interested person ignorant of his true position (p. 122). If the first move is read in this latter, problematic way, then a mistake has already been made, before we even move from the ignorance to equal chance characterizations. We should think of the interests in question as "simply those of the members of the society to whom the principles of justice are to apply" (compare the remarks in the text about people's interests engaging the Pareto preference directly), and Rawls' "reduction of the problem to the case of single person's self-interested choice should arouse our suspicion" (p. 124). (Compare Scanlon's criticisms of the first move with Sandel's remarks about "what really goes on behind the veil of ignorance" and on the way a "voluntarist interpretation of the original position gives way to a cognitive one", in 1982, pp. 122-132.)

The burden of Scanlon's criticism falls not on the assumption of a veil of ignorance, of which he admits there is a valid version, but on the reduction to the point of view of a single rational individual behind that veil. In this respect the view taken here is related to Scanlon's: it exploits the veil of ignorance as a corrective device, but without assuming an exercise of constrained self-interest. The argument here goes further by pointing out the way in which uncertainty aversion can operate, even if we do not revert to the constrained self-interest picture. The influence of attitudes to uncertainty on decision is not tied to evaluations in terms of self-interest, any more than the pointfulness of the Veil of Ignorance is.

9. It may be objected that once we reject the idea of an exercise of constrained self-interest, there is no reason to stop short of full-fledged impartial spectator theory. The argument here is indeed aiming to occupy a middle ground between these options. The role given to the Pareto preference is consistent with the perspective of an impartial spectator. But we can nevertheless use the decision-theoretic connection between uncertainty and distribution to structure the argument to maximin, in the way suggested. This connection can hold even though self-interest is not the relevant motivation. Here I am indebted to objections made by Derek Parfit.

10. Compare the contrast drawn by Rawls in terms of "comprehensive doctrines each with its own conception of the good", in 1993, p. 134-5; compare also "The Idea of an Overlapping Consensus" in that volume.

11. It may be objected that divisions of epistemic labor that do not give each individual's opinion an equal influence on the result are for that reason undemocratic. On this view, individual rights over certain private issues, judicial review, even federal systems that give regions equal influence but not individuals, would be regarded as undemocratic. Hurley 1989, ch. 15, argues that this objection presumes an over-simple conception of democracy, which cannot account for the respects in which certain divisions of epistemic labor are more democratic than others, any more than it can account for the respects in which certain ways of drawing boundaries are more democratic than others.

12. Worth considering in this context are Nussbaum's suggestions about the role of praise and blame in Aristotle's account of the natural animal basis for the development of moral character, and the contrasts she draws with a more Kantian view of responsibility. See Nussbaum 1986, pp. 282-287.
13. Shallow deterrence is normally assumed to be in question when the preventative effects of punishment are assessed in utilitarian terms. But if a system of punishment has deep, long-term influence on social norms and on the formation of characters and desires of persons growing up in society, then these influences may in principle be equally subject to utilitarian calculation.
14. Are side-by-side states in the United States with different sentencing policies good counterexamples, if despite these differences people pay little attention to state boundaries in going about their business? State boundaries might be largely ignored in many aspects of ordinary life for reasons that have little to do with the criminal law, yet these aspects of ordinary life may create occasions for crime. What is the difference between two distinct systems with different sentencing policies, which might provide a controlled experiment, and one inclusive system with what is in effect a sentencing policy with a random element?
15. Many of the points of the preceding two paragraphs paraphrase Andenaes 1974, chapters 1, 2, 4 and 5. Cf. Lacey 1988, pp. 28, 182-3, etc.
16. Here I am indebted to Raymond Geuss.

REFERENCES

- Andenaes, Johannes (1974), *Punishment and Deterrence*. Ann Arbor: University of Michigan Press.
- Bacharach, Michael and Susan Hurley, eds. (1991), *Foundations of Decision Theory: Issues and Advances*. Oxford: Blackwell.
- Barry, Brian (1989), *Theories of Justice*. London: Harvester Wheatsheaf.
- Churchland, Paul M. (1995), *The Engine of Reason, the Seat of the Soul*. Cambridge, Mass.: MIT Press.
- Cohen, G. A. (1989), "On the Currency of Egalitarian Justice," *Ethics* 99 (4), 906-44.
- (1992), "Incentives, Inequality and Community," in *The Tanner Lectures on Human Values*, Vol. 13, Grethe Petersen, ed. Salt Lake City: University of Utah Press, 263-329.
- Condorcet, (1785), "Essai sur l'application de l'analyse a la probabilité des décisions rendues a la pluralité des voix," in *Sur les Elections et Autres Textes*, Olivier de Bernon, ed. Paris: Fayard, 1986.
- Ellsberg, Daniel (1961) "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics*, 643-69.
- Feinberg, Joel (1970), "The Expressive Function of Punishment," in *Doing and Deserving*. Princeton: Princeton University Press, 95-118.
- Gärdenfors, Peter and Nils-Eric Sahlin (1982), "Unreliable Probabilities, Risk Taking, and Decision Making," *Synthese*.

- Griffin, James (199), . Oxford: Oxford University Press.
- Hampton, Jean (1984) "The Moral Education Theory of Punishment," *Philosophy and Public Affairs* 13, 208-38.
- Hurley, S. L. (1989), *Natural Reasons*. New York: Oxford University Press.
- (1993), "Justice Without Constitutive Luck," in *Ethics, Royal Institute of Philosophy Supplement*, Vol. 35, A. Phillips Griffiths, ed. Cambridge: Cambridge University Press, 179-212.
- (1995) "Troubles with Responsibility," *Boston Review* 20 (2), 12-3.
- Hutchins, Edwin (1995), *Cognition in the Wild*. Cambridge: MIT Press.
- Kymlicka, Will (1990), *Contemporary Political Philosophy*. Oxford: Clarendon Press.
- Lacey, Nicola (1988), *State Punishment*. London: Routledge.
- McClelland, Edward F. (1990), *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.
- Mill, John Stuart (1958), *Considerations of Representative Government*. New York: Liberal Arts Press.
- Nussbaum, Martha C. (1986), *The Fragility of Goodness*. Cambridge: Cambridge University Press.
- Raiffa, Howard (1961) "Risk, Ambiguity, and the Savage Axioms: Comment," *Quarterly Journal of Economics*, 690-4.
- Rawls, John (1971), *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.
- (1985) "Justice as Fairness: Political not Metaphysical," *Philosophy and Public Affairs* 14.
- (1993), *Political Liberalism*. New York: Columbia University Press.
- Roemer, John (1985) "Equality of Talent," *Economics and Philosophy* 1, 151-87.
- (1986) "Equality of Resources implies Equality of Welfare," *Quarterly Journal of Economics* CI, 751-84.
- (1987) "Egalitarianism, Responsibility and Information," *Economics and Philosophy* 3, 215-44.
- (1993) "A Pragmatic Theory of Responsibility for the Egalitarian Planner," *Philosophy and Public Affairs* 22.
- (1995) "Equality and Responsibility," *Boston Review* 20 (2), 3-7.
- Ross, Alf (1975), *On Guilt, Responsibility and Punishment*. London: Stevens and Sons.
- Sandel, Michael J. (1982), *Liberalism and the Limits of Justice*. Cambridge: Cambridge University Press.
- Scanlon, T. M. (1982), "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, Amartya Sen and Bernard Williams, eds. Cambridge: Cambridge University Press, 103-28.

--- (1988), "The Significance of Choice," in *The Tanner Lectures on Human Values*, S. MacMurrin, ed. Salt Lake City: University of Utah Press, .

Sen, Amartya K. (1970), *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Strawson, Galen (1986), *Freedom and Belief*. Oxford: Clarendon Press.

Taylor, Charles (1985), *Philosophy and the Human Sciences*. Cambridge: Cambridge University Press.

Waldron, Jeremy (forthcoming) "Deliberation, Disagreement and Voting," *Working Paper for Shell Center for International Human Rights*.