

DECEPTIVE PROPERTIES AND MODELING SOCIAL CHANGE

Brian Epstein

Department of Philosophy, Virginia Tech

Suppose P is an intrinsic property that holds of individuals.¹ For example, let P be a psychological property, such as having a preference or attitude that can plausibly be regarded as intrinsic. Suppose we introduce a new property P' which holds of groups, and define P' to be a combination or aggregate of P , over the members of the group. For example, we might define P' as holding of a group G just in case every member of G has P . If, for example, we take P to be the property *being selfish*, and every member of the legislature has that property, then the legislature as a whole has the property we might call *collectively being selfish*.

Suppose we are not happy the legislature has this property, and we want to change it. We want to enact policies, that is, to make it the case that P' will no longer hold of the legislature. To do this, we put together a mathematical model of P' , so that we can assess what levers are the most effective and inexpensive for changing it, and then put policies in place that employ the best of those levers.

There is an obvious approach to building a model of P' : we consider the factors that affect whether property P holds of individual legislators. Since P' is simply the aggregate of P , the way to change P' is to pick the levers that cause changes in P . In this case, we consider policies that promote altruism in legislators. For instance, we might try

¹ This means that, as applied to individuals, P is globally intrinsic, in the sense in Humberstone (1996), i.e., that it is intrinsic to any entity possessing it. I will distinguish a property such as “intending to do such-and-such” that applies to individuals from the shared intentional property “intending to do such-and-such” that applies to groups.

out a program of civic education, or one in which we bring the Dalai Lama in to give a speech, or one in which we administer electric shocks. The policies that affect P are the ones we choose among, in order to change the group property P'.

My aim in this paper is to show that this reasoning, while natural, is fallacious. Much of social theory is directed toward modeling the properties of groups like the Senate, the Supreme Court, a crowd, or a class of freshman, – i.e., social entities that are made up of individual people. But the vast bulk of models of such properties systematically ignore the fact that many factors can affect the P'-like properties applied to a group even if they have no effect at all on the P-like properties applied to individuals. And if these models are faulty for modeling change in artificial properties like P', which are very simple aggregates, then *a fortiori* they are faulty for more natural social properties of groups. My aim is to take a step toward eliminating this intuitive but mistaken approach, to expand the resources for understanding and effecting social change.

Electoral control

To show how this problem arises in a practical context, I will begin with a particular case in social theory.

One of the key concerns for the design of political systems is to ensure that politicians, such as legislators, act in the interests of their constituents. It is not difficult for social theorists to explain why a legislature does not always do what its constituents want. Politicians are rational people, like the rest of us, and when their interests diverge from their constituents' interests, their choices will be affected. The idealist wonders why the choices of the legislature deviate from the interests of the constituents, but the

political scientist takes the converse to be the puzzle: why does the legislature do what is in the constituents' interests at all? How does an electorate manage to control the actions of the legislature?

This is known in economics as a "principal-agent" problem. A principal (the electorate) hires agents (the legislature) to perform some task that serves the principal's interests. How does the principal get the agent to act in alignment with the principal?

From the perspective of understanding social change, this is an interesting phenomenon at a couple of levels. How the legislature thinks, behaves, and votes are social properties of a group, so they serve as examples of the kind of properties that social theory is in the business of modeling. These particular properties, however, are not only good as examples of group properties, but are important in their own right, if we are to figure out how to implement social change.

An obvious way for the electorate to get the legislature to do its bidding is to set up a system of rewards or punishments. These, however, are difficult to implement for legislators. The big problem with direct incentives, monetary ones in particular, is that they tend to stack the outcomes in alignment with the interests of the wealthy, not with those of the electorate as a whole. Corporal punishment or solitary confinement do not seem to be workable either.

Fortunately, in electoral systems there is a second lever for controlling the behavior of legislators. If the electorate is unhappy with the way the legislature is voting, it throws them out of office. If elections are extremely infrequent, or if there is a one-term limit for legislators, then this mechanism does not work. In a political system with frequent elections, however, the replacement of misbehaving elected officials can be a

powerful force for ensuring the conformance of the vote of the legislature with the preferences of the population.

The modeling of electoral control

A number of mathematical models have been developed to treat these mechanisms for electoral control. The predominant models in the field are descendents of those developed by Robert Barro and by John Ferejohn of how equilibrium is reached in a strategic game between the politicians and the electorate.²

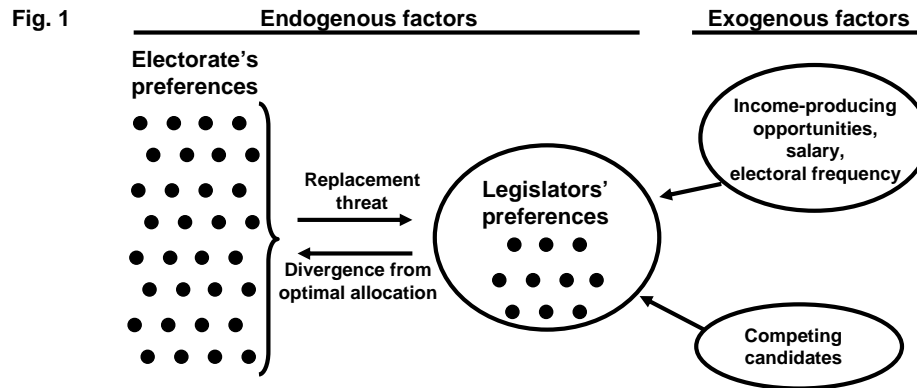
To model the tradeoff between the interests of politicians and the electorate, they consider the case of competing financial interests. They begin with the assumption that a politician can extract “political income” from being in office, in the form of bribes, kickbacks, and so on. These payments change the allocation of social resources, away from the optimal level favored by the electorate. If an officeholder extracts the maximum political income in a term, then the electorate will choose not to re-elect her. But if the officeholder extracts no political income, then even though she is re-elected, she has not extracted the potential benefits of being in office, so she might as well have extracted the maximum and failed to be re-elected.

This sets up the parameters for a strategic game between the parties. Voters maximize their well-being, subject to the constraint that legislators act in their rational self-interest. The amount of “electoral control” that the electorate can exercise, in virtue of the threat of non-reelection, is the difference between the equilibrium level this implies and the maximum that the politician could otherwise extract. In these models, the level of electoral control ends up depending on such factors as the overall value of the office,

² Barro (1973), Ferejohn (1986).

political salary, electoral frequency, and the number of competitors for office.

Figure 1 is a simple representation of the variables treated in the Barro model and its descendents:



These models incorporate a variety of influences on the legislature, and interactions the legislature has with the population. There are some exogenous variables in these models, such as the political salary, the electoral frequency, and income-producing opportunities. Also usually treated as exogenous is the amount of competition for office, or the replacement pool of candidates. All of these exert influence on the incentives and hence the decision-making of the legislators in office. The endogenous part of the models is the strategic interaction between the population and the legislators. The legislators exert influence on the population, through the divergence between their votes on how to allocate resources and how the population optimally wants them to do so, and the population exerts influence on the legislators by threatening to replace them.

There are a number of interesting challenges to building a good model of this strategic interaction. One is the problem of the final term or "last period" a politician is in office. Supposing that politicians retire at a certain point, on this explanation there is no pressure for politicians to conform to the interests of the electorate in that period. Thus their last period is not subject to electoral control. This means that it is rational for

the electorate to vote politicians out in the term prior to their retirement term. Yet that makes the previous term the politicians' expected last period, which generates the same problem. The lack of electoral control in the last period thus cascades back to the penultimate period, and the antepenultimate period, and so on.

This sort of problem is familiar in game theory, and political scientists have built increasingly sophisticated models, designed to explain the actual conformance and divergence of legislative behavior with the will of the electorate, and also designed to provide policymakers with a set of levers to manipulate, to improve this conformance.

The problem I wish to discuss is that there is a whole class of levers and potential interventions that these models neglect entirely. The models focus on the strategic game between the legislators and the populace. But, as I will argue, there is a wide range of policy options for affecting the vote of the legislature that have no connection – potentially even no causal connection – to the legislators at all. The reason for this gap is that modelers seem to have been deceived by the properties they model.

The deceptive extrinsic

Intuitively, an intrinsic property is “a property that a thing has (or lacks) regardless of what may be going on outside of itself.”³ The mass and charge of an object are among its intrinsic properties, while *being a husband* and *being ten miles from Berkeley* are extrinsic. It is sometimes difficult, however, to distinguish which properties are extrinsic: only in the seventeenth century did we learn that *being heavy* is an extrinsic property; some people think that *being red* is an intrinsic property, while others believe it to be extrinsic; and in the philosophy of mind, there is controversy as to whether mental

³ Yablo (1999)

properties such as *having the thought that the sky is blue* is intrinsic or extrinsic. (Incidentally, content externalism is an issue I will put to the side here: for these purposes I will assume internalism.)

A striking and counterintuitive point that has come to light in recent years is that for certain objects, their identity properties are extrinsic to them. Intuitively, it would seem that for any object *O*, the property *being O* is intrinsic to that object, but this is not so. Stephen Yablo, for instance, has pointed this out with an argument having the following structure:⁴

(1) For an organism to be a member of a particular species depends on the evolutionary history of that organism. That is, species membership is an extrinsic property of an organism. For Rocky to be a squirrel, for instance, depends on Rocky's evolutionary history.⁵

(2) The property *being of such-and-such a species* is plausibly an essential property of any organism that has the property. If it is true that Rocky is a squirrel, for instance, then *being a squirrel* is among Rocky's essential properties.⁶

Therefore,

(3) The identity property of an organism is an extrinsic but essential property of any organism that has that property. For instance, the property *being Rocky* is an extrinsic property of Rocky.

The application of this to the present context comes in two pieces: first, the extrinsicness of certain identity properties makes other properties, even ones that at first

⁴ Yablo (1999)

⁵ Hull (1978), Boyd (1999), Ereshefsky and Matthen (2005), among others.

⁶ Kripke (1980), Soames (2002), among others.

blush seem intrinsic, inherit that extrinsicness. And second, social groups are among the objects that typically have extrinsic identity conditions.

Inheriting extrinsicness happens for any property that applies only to objects having extrinsic essential properties. To give a somewhat artificial example, consider properties that it is essential to be human in order to possess. It is plausible, for instance, that only humans can possess certain moral or religious properties, such as *being evil* or *being sinful*. But for a given person to be sinful may only be a matter of that person being in a certain mental state. Maybe it is a matter of having a disposition of some sort, such as being lustful in one's metaphorical heart. Given that assessing evil is just a matter of examining a person's mental state, it is natural to assume that *being evil* is an intrinsic property. If holding the property genuinely depends on being human, however, then it is in fact extrinsic: it depends on all the historical factors that *being human* depends on. The property *being evil* looks intrinsic, but in fact is what we might call a "deceptive extrinsic" property.

As soon as it is seen that objects can have extrinsic identity conditions, it is clear that this is a feature of many social entities. Groups such as the Senate, the American citizenry, the Supreme Court, and the Berkeley freshman class – are individuated by factors that are not local to the members of the groups. (One common way of putting this is that a social property like *being a Senator* does not supervene locally on individualistic properties, but the point is clear even without using the machinery of supervenience.) Many such groups have properties that only apply to them, and thus there straightforwardly are many social properties that are deceptively extrinsic, i.e., properties that would be intrinsic but for the extrinsicness of the identity property of the only group

to which they apply. The property *having voted for cloture* or the property *having elected the members of the Committee on Foreign Relations*, for instance, is a property that only the Senate (or a limited number of other legislative bodies) can have. The possession of this property by an object requires the object to be a group with a certain externally determined status.

Having voted for cloture is an extrinsic property because it only applies to objects whose identity property is extrinsic. The cases that I began with, the P' cases such as *collectively being selfish*, are somewhat different. These are genuinely intrinsic properties of a group. They are determined entirely by the intrinsic properties of the members of the groups to which they apply, and they apply to groups that either do or do not have essential extrinsic properties. However, a similar kind of deceptiveness can arise with these properties as well. The reason is that in social theory, we almost always model these properties in their application to a specific group or type of group. For instance, we do not simply intend to model the property *collectively being selfish* as applied to any group whatsoever, but rather the property as applied to the legislature, because we are interested in policies that will change its selfishness in particular. When the object that a property applied to has an essential extrinsic property, in effect the property being modeled itself becomes extrinsic. This point is even clearer if we cast what is being modeled not as a property being applied to a particular social group, but rather as a property that applies to circumstances as a whole. What we want to model, and what we want to change with our policies, is the property of the world: *being such that the legislature is selfish*.

When we analyze, model, or manipulate an extrinsic social property, its

dependence on external factors is crucial. Consider the property *having committed perjury*, versus the property *having lied*. In both cases, one comes to possess the property by asserting falsehoods. But to be a perjurer, one also needs to have a particular status, namely, being bound by a legal oath, such as in a trial or deposition. To be a liar arguably involves only intrinsic requirements, but to be a perjurer those intrinsic requirements are supplemented by an extrinsic social status. Given that *being a perjurer* is extrinsic, one way of changing the property is to change only the external factors determining whether the property holds. For instance, we can eliminate perjury entirely by eliminate all legal oaths, without eliminating a single lie. If we assume that the only way of reducing perjury is to reduce lies, then we conflate the extrinsic property of interest with the intrinsic property that is just one component of it. A good model of the levers affecting perjury might be very different from a good model of the levers affecting lies.

The same point applies to the modeling of a property such as *collectively being selfish* to a group having an essential extrinsic property. Just as with an extrinsic property like *being a perjurer*, there are two ways of changing such a property in the world. We can either change the psychological properties of particular legislators, or we can change the property *being the legislature*. For instance, one way of changing the circumstance so that *being such that the legislature is selfish* does not hold is to dissolve the legislature altogether. Because *being the legislature* is an extrinsic property of the legislature, this can be done without interacting with any individual legislators whatsoever.

The blind spot in models of electoral control

To illustrate how missing this point generates a significant blind spot in prevailing

models of electoral control, consider how these models apply to a troublesome case: the endemic problem of political corruption in Italy. Italy is consistently ranked near the bottom among European countries on international corruption indices. This is a longstanding problem, but little that the Italians do seems to make a difference. Not only are the rules for governance in Italy comparable to other European countries, but the enforcement of those laws is comparable. And there is little evidence that other countries have any more effective systems for punishing corrupt officials than Italy does, or greater disincentives for corrupt behavior on the part of elected officials. The key problem with the Italian system, instead, seems to be that the ability to replace members of parliament is impaired. One political observer recently said that Italian politics is like a game of musical chairs, played over and over again, but without removing any chairs.

Consider how prevailing models of electoral control would model this. They would explain the persistence of corruption in Italy as arising from the fact that since legislators know that they will not be replaced for corruption, they have no incentive to compromise in following their own interests. The legislators recognize that the failure of replacement mechanisms effectively puts them in the same situation as if they were in the “last period,” so they have no incentive to hold back from enriching themselves.

As I mentioned above, these models have an important virtue that models of group properties often do not: they include both endogenous and exogenous factors apart from the legislature, which cause changes in the preferences of the legislature. Notice, however, how the external factors are employed in the model. The property being modeled is treated as fully determined by the intrinsic properties of individuals in the legislature. All the other factors in the model are incorporated because they have a causal

influence on those intrinsic properties. The exogenous factors have a unidirectional influence, and the endogenous factors interact bidirectionally with the legislators. But the effect of any factor is only considered in its causal effect on the legislators. Thus factors that are not causally connected to actual legislators are neglected altogether. The inability of the electorate to succeed at replacing legislators does figure into the model, but it only does so in one way: in as much as it affects the strategic game between the electorate and the legislators.

In the case of Italy, however, there is reason to think that this strategic game, and the setting of incentives altogether, is not the critical factor. The downside of losing office is one of many factors that determine the incentives of the legislators. But lots of factors determining legislator incentives have been manipulated in Italy, to little effect. The resilience of corruption in the face of repeated changes to enforcement mechanisms suggests that legislators may just be insensitive to manipulations of their incentives. If the likelihood of receiving a fine or prison sentence is not adequate to tilt the legislators' incentives enough to reduce corruption, then why should the threat of losing office be an effective lever?

Putting incentives aside, there is a straightforward alternative explanation of what damage the "musical chairs" electoral system does in Italy. The most obvious virtue of electoral replacement has nothing to do with inducing changes in behavior through making it rational for legislators to conform their votes to the preferences of the electorate. Instead, it is simply that without effective electoral replacement, bad or nonconforming legislators have no way of being replaced by good or conforming candidates. Effective electoral replacement has the virtue that it takes advantage of the

heterogeneity of candidates, appropriately changing the extension of the property *being a legislator* when there is nonconformance.

Replacement can change the attitudes or behavior of the legislature even if incentives are left entirely unchanged. And in fact, given that these properties are extrinsic, they can be changed even if the preferences, beliefs, and behavior of all actual and potential legislators are left entirely unchanged. Suppose we have a pool of 300 heterogeneous individuals, 100 of whom are legislators, and 200 of whom are not. We can leave the intrinsic properties of all 300 people entirely unchanged, and yet change the preferences of the legislature. This is done simply by changing the status of who is a legislator, which can be done by changing factors entirely disconnected from the candidates themselves.

To illustrate this differently, let us stipulate that the incentives of legislators are the same throughout their tenures as if they were in the “last period,” to wipe out any effects of elections on incentives. Suppose that there is no punishment at all for legislators who are found to have enriched themselves. Legislators can be terminated, but when they are, they receive severance payments that make them whole. From the legislator’s perspective, she is indifferent to staying in office or being tossed out. Nonetheless, even having negated the influence of incentives, the electorate has significant power to control the behavior of the legislature. If we only presume heterogeneity in the character or preferences of the candidate population, or some continuity in the way legislators will vote over time, then the electorate can change the legislative outcomes by replacing legislators with others, even while their incentives are unchanged.

This mechanism for the electorate's exercising control is arguably the dominant one in most actual political systems. It is probably not too far from the truth to assume that voting patterns are fixed for any given candidate long before she becomes a legislator, and that they can only be changed with the greatest efforts. Electoral control is exercised most commonly not by setting up systems in which candidates from a homogeneous pool will have the appropriate incentives to be agents, but by filtering a heterogeneous pool of candidates for the preferences that already match those of the electorate.

It is not a criticism of models like Barro's and Ferejohn's, that they neglect this. Models are simplifications, and no single model can be faulted for choosing to focus on one set of levers or explanatory factors above another. It becomes a problem, however, when all models are constructed with a systematic bias in favor of local factors as the determining ones, and ignoring the nonlocal factors except in their causal influence on local ones.

As I have mentioned, models of electoral control are already broader than many models of the properties of social groups, in that they take into account properties outside of the individuals in the group itself as endogenous variables. Most models of the properties of groups model only the intrinsic properties of members of the group, and include a few fixed environmental parameters as exogenous variables. These electoral control models seem not to be so blinkered: they do include the properties of the electorate endogenously in the model. But even these broad-minded models do not escape the implicit assumption that an intrinsic property of a group is fully determined by the modeling the influences on the members of the group.

To return to the modeling of *collectively being selfish* as applied to the legislature: the changes in the property being modeled do not need to arise from changes in the selfishness of any individual, but can instead arise from the rejiggering of the legislature. To approach the modeling of P' by focusing on P is potentially misleading, leaving out a range of potentially relevant factors by fiat. This potentially takes a significant toll on effecting social change. The best levers for change can be lost with models that are deceived into too limited a view of the determinants of the properties they model. If the goal is to attack social problems, these models may inadvertently lead us to bring the knife along to the fight, but to leave the gun in the bedside drawer.

Deceptive properties and the analysis of collective intentionality

Let me conclude by commenting on a connection between the issues I have discussed here and some more familiar issues in recent discussions of collective intentionality.

Much work on collective intentionality has been to attack what we might call “the fallacy of decomposition.” In Michael Bratman’s theory of shared intention, for instance, for a group to have the intention to paint a house, it is not sufficient for the individuals each to have the intention to paint the house.⁷ It is an error, that is, to infer that the possession of the intentional properties of groups is simply an aggregation of the same intentional properties in the individual members of the groups. Rather, there need to be further conditions on the members of the group, such interlocking intentions.

A related area attacking such a fallacy is the work many people are doing on problems of judgment aggregation. This work is particularly interested in cases in which

⁷ Bratman (1993)

group judgments may be at odds with the judgments of all their members. There are tradeoffs that need to be made in determining what property deserves to be called the judgment of a group. Philip Pettit has argued that on this basis groups can be understood as having a “mind of their own,” where the intentional properties of a group supervene on but do not have an orderly correspondence to the intentional properties of the group members.⁸

The foregoing discussion, however, implies that there is another mechanism for the intentional properties of groups to be at odds with the intentional properties of individuals. Like the selfish legislature, a group can have a mind of its own simply based on what the intentional property is a property of. In his discussion of the mental lives of groups, Pettit cites work by himself and by Gregory Currie, in which he notes that his claims about the independent mental life of groups is compatible with the supervenience of the group properties on individualistic ones.⁹ Ironically, in those very papers, both he and Currie stress that while *global supervenience* is true,¹⁰ *local supervenience* is false. Yet in treating group intentional properties in such models as shared intention and judgment aggregation models, it is implicitly assumed that the group intentional properties *are* locally supervenient. The present considerations suggest that for many group properties being analyzed and modeled in the way judgment aggregation is, this assumption may well be mistaken.

⁸ Pettit (2003)

⁹ Macdonald and Pettit (1981), Currie (1984), Pettit (2003).

¹⁰ This is a point I dispute elsewhere (Epstein (2007)), but resolving this is unnecessary for the present discussion.

REFERENCES

- Barro, R. (1973). "The Control of Politicians: An Economic Model." *Public Choice* 14: 19-42.
- Boyd, R. (1999). "Homeostasis, Species, and Higher Taxa." (In *Species: New Interdisciplinary Essays*. R. Wilson, Ed. (pp. 141-185). Cambridge: MIT Press.)
- Bratman, M. (1993). "Shared Intention." *Ethics* 104: 97-113.
- Currie, G. (1984). "Individualism and Global Supervenience." *British Journal for the Philosophy of Science* 35: 345-58.
- Epstein, B. (2007). "Ontological Individualism Reconsidered." *Synthese* (In press; available online).
- Ereshefsky, M. and M. Matthen. (2005). "Taxonomy, Polymorphism, and History: An Introduction to Population Structure Theory." *Philosophy of Science* 72: 1-21.
- Ferejohn, J. (1986). "Incumbent Performance and Electoral Control." *Public Choice* 50(1-3): 5-25.
- Hull, D. (1978). "A Matter of Individuality." *Philosophy of Science* 45: 335-360.
- Humberstone, I. L. (1996). "Intrinsic/Extrinsic." *Synthese* 108(2): 205-267.
- Kripke, S. (1980). *Naming and Necessity*. (Cambridge: Harvard University Press.)
- Macdonald, G. and P. Pettit. (1981). *Semantics and Social Science*. (London: Routledge & Kegan Paul.)
- Pettit, P. (2003). "Groups with Minds of Their Own." (In *Socializing Metaphysics*. F.

Schmitt, Ed. Lanham, MD: Rowman and Littlefield.)

Soames, S. (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and*

Necessity. (Oxford: Oxford University Press.)

Yablo, S. (1999). "Intrinsicity." *Philosophical Topics* 26: 479-505.