Invisible-Hand Explanations

By Robert Nozick*

In Nozick (1974), I described how, if people entered into mutual protection agreements and firms offered buyers protective services, a dominant protection agency would arise by legitimate steps, and this would constitute at least an ultra-minimal state. No one need have intended to produce a state. A pattern or institutional structure that apparently only could arise by conscious design instead can originate or be maintained through the interactions of agents having no such overall pattern in mind. Following Adam Smith, I termed such a process or explanation an *invisible-hand* process or explanation and offered a list of examples to make the phenomenon salient. These included evolutionary explanations of the traits of organisms and populations, microeconomic explanations of equilibria, Carl Menger's explanation of how a medium of exchange arises, and Thomas Schelling's model of residential segregation. (Edna Ullmann-Margalit [1978] is a later attempt to define the concept.) Two types of processes seemed important: filtering processes wherein some filter eliminates all entities not fitting a certain pattern, and equilibrium processes wherein each component part adjusts to local conditions, changing the local environments of others close by, so the sum of the local adjustments realizes a pattern. The pattern produced by the adjustments of some entities might itself constitute a filter another faces. The opposite kind of explanation, wherein an apparently unintended, accidental, or unrelated set of events is shown to result from intentional design, I termed a *hidden-hand explanation*. The notion of invisible-hand explanation is descriptive, not normative. Not every pattern that arises by an invisible-hand process is desirable, and something that can arise by an invisible-hand process might better arise or be maintained through conscious intervention.

Economics typically explains patterns in terms of the actions of rational agents. However, a disaggregated theory of the agent herself, wherein patterns that seem to indicate a central and unified directing agent are instead explained as the result of smaller, non-agent entities interacting, also might count as an invisible-hand explanation.[1] The definitional details of what counts as "invisible hand" are less interesting than the particular theories.[2]

Time preference seems susceptible to evolutionary explanation (see Nozick, 1977; and Nozick, 1993 pp. 14–15). The future is uncertain, an organism may not survive to reap an anticipated reward, or the world might not present it. Innate time preference

*Department of Philosophy, Harvard University, Cambridge, MA 02138. This essay is dedicated to the memory of Raymond Lubitz, 1937–1984, A.B. Columbia (1959), B. Phil. Oxford (1961), Ph.D. in Economics, Harvard University (1967), Assistant and Associate Professor of Economics, Columbia University (1967–1973), Federal Reserve Board, Washington, DC (1973–1984), and Chief of its World Payments Economic Activities Section, Division of International Finance; coauthor of Kenen and Lubitz (1971).

[1]Daniel Dennett (1991) proposes such a disaggregated theory of the self. Question: what decentralized competing processes *within* an individual would give rise to a (relatively) coherent decision-maker?

[2]For instance, a theory would be interesting if it showed that, although everyone *was* aiming at a pattern, either their actions animated by that aim were not what produced the pattern, or if they did, that the pattern did not arise by the route everyone imagined—it was a side effect of their envisioned plans.

may be evolution's way to instill in creatures incapable of explicit probabilistic calculations a mechanism having roughly the same effect, approximating what such calculations would have yielded with regard to rewards affecting inclusive fitness; such time preference may have been selected for. Consider, then, beings with the cognitive apparatus to take explicit account of such uncertainties, who explicitly perform a probabilistic discounting of the future. If already installed in humans is an innate time preference—evolution's attempt to perform that probabilistic discounting for us—and if what we explicitly discount in our probabilistic calculations is the (already discounted through time preference) present value of the future reward, then what takes place will be a *double discounting*. Isn't that too much?

Next, consider wealth-maximization or the weaker assumption that people are seriously concerned with wealth. A widespread phenomenon across societies (though not within Western industrialized societies in the last 150 years) is that wealthy people tend to have more children. (Gary Becker [1981 p. 102] cites supporting literature.) Suppose that, ceteris paribus, people with a strong desire for wealth tend to amass more; it is more likely that they will. If there had been a genetically based heritable psychological predisposition to be (more) concerned with wealth—I do not claim this as anything more than a possibility—then that would have been selected for; the percentage with that heritable desire would increase over time. This would provide an evolutionary explanation of, if not wealth maximization, a widespread strong desire for wealth (see Nozick, 1993 pp. 126–27).

Evolutionary explanations also can be brought to bear within philosophy to explain a priori knowledge of apparently necessary truths and to explain the intractability of certain philosophical problems (see Nozick, 1993 Ch. 4). Traditional philosophical doctrine attributes to individuals a faculty of a priori knowledge, enabling them to know independently of experience that certain things must hold true, that they hold true in all possible worlds. It is implausible that evolutionary processes, keyed to the actual world, would instill any such completely general faculty within us.

Yet certain propositions do seem self-evident, and it is difficult to think of ways they might be false. A certain proposition's *seeming* self-evidently true to us might have been selected for, if it does hold true (at least approximately) and if acting upon a belief in this does, in general, enhance fitness. That factual, contingent truth would come to seem *more* than just factual, through evolutionary selection via the "Baldwin effect": those to whose "wiring" a connection or proposition seems closer to evident learn it faster and gain a selective advantage; they leave offspring distributed around their own speed of learning until, over generations, all find it self-evident. If, frequently enough, samples of a certain sort resembled their populations, then generalizing from samples to population, or to the next encountered member, would frequently yield truths, and those to whom such inferences seemed obvious and self-evident would frequently arrive at those truths.

Rationality itself might be an evolutionary adaptation. Evolution phylogenetically instills in us information about, and patterns of behavior suitable to, stable facts of our evolutionary past. Evolution utilizes and builds mechanisms around constant and stable environmental features (e.g., gravitational force is utilized in the working of some physiological processes, which were designed to utilize and function *in tandem with* steady gravity, not to duplicate separately what gravity already does).

Some obdurate philosophical problems (e.g., justifying induction, or our belief in the existence of other minds or in an external world) might mark stable facts about humans' past environment that evolution has built into us as *assumptions*, marking facts to work in tandem with. All human beings heretofore have been born in environments surrounded by other people with minds similar to their own, in an independently existing "external world," one whose objects continued on trajectories (or in place) even when unobserved, a world in

which certain kinds of generalization or extrapolation from past experience led to further truths. Those who failed to learn this quickly left fewer similarly uncomprehending descendants. Rationality's function was not to *justify* these assumptions but to utilize them.

Hidden-hand explanations, the opposite of invisible-hand ones, tend toward ruling-class (or, more extremely, conspiracy) theories. What a ruling class aims at and produces or maintains is not given an invisible-hand explanation. However, the existence of a ruling class might itself be given such an explanation, if it did not arise as the result of some individual's or group's actions intending to bring this about.

Here is a sketch of how this might occur. Start with a society containing no ruling class, where the most powerful and wealthy individuals want their children and grandchildren to be equally or more advantaged and so place them in environments (schools, vacation places) that make more likely their children's marrying similarly advantaged people. Marriages forge alliances of mutual interest, making more likely the sharing of information and coordinated activities for mutual benefit. Allies and employees will tend to be recruited from similar schools and social networks, because their families are directly known, or because the similar molding of their values, tastes, and modes of behavior makes them easier to work with, more predictable, more congenial, less likely to create conflict. Directors of companies will be recruited from among similar persons already successful elsewhere; studies of boards of directors would show similar social backgrounds and much interlocking.

Matters of mutual interest are discussed, including public matters; sometimes joint representation is made to government officials about matters of mutual concern. As issues become complex, or the polity becomes widespread, organizations are started to think through these issues together and to interact with the government officials (or potential ones) who might significantly affect them. Thus might arise a pattern of wealthy and powerful individuals associating in social, business, and political life. How much *success* and *coordination* in the determination of (which?) results is needed for this to constitute, technically, a ruling class? The coordinating organizations might be started, maintained and participated in without the *aim* of "serving the interests of the ruling class," so a ruling class might arise through an invisible-hand process, even if later it consciously maintains itself.

Not just equilibria within markets, but the very existence of markets in the West is largely the product of invisible-hand processes. People aimed to extend particular markets in one or another direction, but "the market" developed bit by bit, unintended. (Even after an overall conception did arise, the extension of markets rarely depended upon economists who had mastered that general conception.) Now, however, there are self-conscious efforts to establish markets and a market society where none had been. If successful, the arising there of a market society will not have an invisible-hand explanation, but particular equilibria within the markets will. Will the new markets' achieving certain overall patterns have an invisible-hand explanation, when those markets were instituted *in order* to achieve just such patterns? (And what of our markets, if they continue to be maintained in part because they are perceived to yield that pattern?) And what shall we say of new institutions, not imitating anything existing before, that are designed and instituted to achieve certain patterns by so structuring incentives that people interacting will produce that pattern? The pattern is invisible to those within the institution but not to its designers.

Here is a suggestion of an institution, call it the *help chain*. With significant publicity and moral suasion, the government institutes a system of help-vouchers, distributing Y hours of help vouchers to every family whose yearly income is below X. A person with such a voucher can request teaching, advice, or help of any individual, and if that individual agrees and delivers it, he receives help-vouchers for that time expended. These vouchers he then can use himself, asking another person for help for himself or for any designated individual. Each person ap-

proached knows that if he agrees, he too will receive a voucher. Unwillingness to ask is reduced by knowing that the other will receive a useful voucher in return; willingness to agree may similarly be aided. Each year, there is a fresh infusion of help-vouchers, starting at the bottom of the income scale, and "trickling up" through voluntary interactions. What new patterns will result?

The standard economist's invisible-hand explanation involves individual agents who choose rationally. (Notice that a theory of irrational behavior also might be specific enough to explain patterns arising from the interaction of individuals behaving in that predictably irrational fashion.) However, the principle of rational decision need not be the principle of maximizing expected utility. In Nozick (1993), I propose a rule of maximizing decision-value, where this is a weighted sum of causally expected utility, evidentially expected utility, and symbolic utility.[3] This rule then is applied to Newcomb's problem and to the prisoner's dilemma, with new results (see Nozick, 1993 Ch. 2). (Newcomb's problem was first presented and discussed in Nozick [1969]; Richmond Campbell and Lanning Sowden [1985] contains articles on this problem plus a bibliography.) New patterns can be explained in invisible-hand fashion as the result of the interaction of agents whose behavior conforms to this broader decision rule.

One also might impose more stringent conditions on preference in addition to the usual structural conditions S (e.g., the von Neumann-Morgenstern conditions). In Nozick (1993), I propose that these additional conditions include the following (which themselves then are also added to the set S): the person prefers satisfying the conditions S to not satisfying them; the person prefers (ceterus paribus) the means and preconditions to satisfying the conditions S; the person prefers having her first- and second-order preferences cohere; the person prefers that the preconditions for making preferential choice obtain, and that the capacities for making and effecting preferential choice not be interfered with. (There are more complicated additional conditions.) When a person's preferences satisfy these (and similar) structural conditions, I say that her preferences are *rationally coherent*.

A plausible view holds that the rationality of a belief depends upon the nature of the process that actually gave rise to (or maintains) it. Simplifying greatly, a belief is rational if it arose by a process that reliably yields true beliefs.[4] Can we demarcate a rational preference as one given rise to by a process that reliably produces ____ preferences? What is to fill in the blank?

One can bootstrap by using the additional structure conditions. A particular preference is rational only if it actually was generated by a process that reliably yields rationally coherent preferences. This requires more than that the preference itself satisfy the additional structural conditions, for the processes that human beings actually can use reliably to yield coherent preferences form a restricted class; and it may be impossible to generate a particular preference by any process in this class, even though it itself does not violate the structural conditions. Given interacting individuals with such stringently rational preferences, some further institutions or patterns might then be explainable.

[3]The evidentially expected utility of an action is the weighted sum of the utilities of its (exclusive) possible outcomes, weighted by their conditional probabilities given the actions. The causally expected utility of an act replaces these conditional probabilities by probabilities (of outcomes on actions) reflecting some direct causal influence.

[4]Notice that, on this view, decision theory is not a theory of rational action, but of best or optimal action. An action would be rational if it were given rise to by a process that reliably yields optimal or maximizing actions. However, a person might happen to stumble confusedly upon doing such a maximizing action. In performing it, he would not be acting rationally, for his action would not be generated by a process that reliably produces optimal actions.

What are the limits of invisible-hand explanations? Many enduring patterns of behavior can be seen as maintained rigidly in the space left by the jigsaw puzzle of other people's actions, where the shape of each of those other pieces is similarly maintained by its surrounding pieces. Are there kinds of institutions or patterns that, in principle, cannot be given an invisible-hand explanation? (Consider written constitutions.) Are there any social structures that could not have arisen by an invisible-hand process, or be maintained by one?[5] If so, is there an illuminating general description of what must evade invisible-hand explanation?

[5]Invisible-hand explanations need not be a subclass of methodological individualist ones. Suppose that some pattern arises at random in particular societies, and also that there exists an irreducible filter (not susceptible to individualist explanation) that eliminates all societies that do not fit that pattern. Than there would be an invisible-hand (but not an individualist) explanation of why all societies fit that pattern (see Nozick, 1974 p. 22). (I sharpen and discuss the notion of methodological individualism in Nozick [1977].)

In his Ely lecture, Kenneth Arrow (1994) refers to common information in the public domain as raising problems for methodological individualist explanation, but it is not evident why either the existence or the consequences of such information (or of *books*!) must elude such explanations.

REFERENCES

Arrow, Kenneth J. "Methodological Individualism and Social Knowledge." *American Economic Review*, May 1994 (*Papers and Proceedings*), *84*(2), pp. 1–9.

Becker, Gary. *A treatise on the family*. Cambridge, MA: Harvard University Press, 1981.

Campbell, Richmond and Sowden, Lanning. *Paradoxes of rationality and cooperation: Prisoner's dilemma and Newcomb's problem*. Vancouver: University of British Columbia Press, 1985.

Dennett, Daniel. *Consciousness explained*. Boston: Little, Brown, 1991.

Kenen, Peter B. and Lubitz, Raymond. *International economics*, 3rd Ed. Englewood Cliffs, NJ: Prentice-Hall, 1971.

Nozick, Robert. "Newcomb's Problem and Two Principles of Choice," in N. Rescher, A. R. Anderson, P. Benacerraf, A. Grunbaum, G. J. Massey, and R. Rudner, eds., *Essays in honor of C. G. Hempel*. Dordrecht: Reidel, 1969, pp. 114–46.

_____. *Anarchy, state, and utopia*. New York: Basic Books, 1974.

_____. "On Austrian Methodology." *Synthese*, November 1977, *36*(3), pp. 353–92.

_____. *The nature of rationality*. Princeton, NJ: Princeton University Press, 1993.

Ullmann-Margalit, Edna. "Invisible-Hand Explanations." *Synthese*, October 1978 *39*(2), pp. 263–91.