

- Hookway, C. (1988). *Quine: Language, Experience and Reality*. Stanford, CA: Stanford University Press.
- Kripke, S. (1982). *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.
- Lewis, D. (1974). Radical interpretation. *Synthese* 27: 331–344.
- Massey, G. (1992). The indeterminacy of translation: A study in philosophical exegesis. *Philosophical Topics* 20: 317–345.
- Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.
- Quine, W. V. (1981). *Theories and Things*. Cambridge, MA: Harvard University Press.
- Tarski, A. (1949). The semantic conception of truth. In H. Feigl and W. Sellars, Eds., *Readings in Philosophical Analysis*. New York: Appleton Century Crofts, pp. 52–84.

Rational Agency

In philosophy of mind, rationality is conceived of as a coherence requirement on personal identity: roughly, “No rationality, no agent.” The agent must have a means-ends competence to fit its actions or decisions, according to its beliefs or knowledge-representation, to its desires or goal-structure. That agents possess such rationality is more than an empirical hypothesis; for instance, as a putative set of beliefs, desires, and decisions accumulated inconsistencies, the set would cease even to qualify as containing beliefs, etc., and disintegrate into a mere set of sentences. This agent-constitutive rationality is distinguished from more stringent normative rationality standards, for agents can and often do possess cognitive systems that fall short of epistemic uncriticizability (e.g., with respect to perfect consistency) without thereby ceasing to constitute agents.

Standard philosophical conceptions of rationality derive from models of the rational agent in microeconomic, game, and decision theory earlier this century (e.g., Von Neumann and Morgenstern 1944; Hempel 1965). The underlying idealization is that the agent, given its belief-desire system, *optimizes* its choices. While this optimization model was proposed as either a normative standard or an empirically predictive account (or both), the philosophical model concerns the idea that we cannot even make sense of agents that depart from such optimality. Related ideal-agent concepts can be discerned in principles of charity for RADICAL INTERPRETATION of human behavior of W. V. Quine (1960) and of Donald Davidson (1980), and in standard epistemic logic (Hintikka 1962). To accomplish this perfection of appropriate decisions in turn would require vast inferential insight: for example, the ideal agent must possess a deductive competence that includes a capacity to identify and eliminate any and all inconsistencies arising in its cognitive system.

While such LOGICAL OMNISCIENCE might appropriately characterize a deity, *prima facie* it seems at odds with the most basic law of human psychology, that we are finite entities. A wide range of experimental studies since the 1970s indicate interesting and persistent patterns of our departures from ideal logician (Tversky and Kahneman 1974), for instance in harboring inconsistent preferences. A more extreme departure from reality is that for such ideal agents, major portions of the deductive sciences would be trivial (e.g., the role of the discovery of the semantic and set-theoretic paradoxes in the

development of logic in this century would then cease even to be intelligible). For a COMPUTATIONAL THEORY OF MIND, where the agent’s deductive competence must be represented as a finite algorithm, the ideal agent would in fact have to violate Church’s undecidability theorem for first-order logic (Cherniak 1986).

The agent-idealizations—within the limits of their applicability—of course have served very successfully as simplified approximations in economic, game, and decision theory. Nonetheless, a sense of their psychological unreality has motivated two types of subsequent theorizing. One type reinforces an eliminativist impulse, that the whole framework of intentional psychology—with rationality at its core—ought to be cleared away as prescientific pseudoscientific theory (see ELIMINATIVE MATERIALISM); a related response is a quasi-eliminativist instrumentalism (e.g., Dennett 1978), where the agent’s cognitive system and its rationality diminish to no more than convenient (but impossible) fictions of the theoretician that may help in predicting agent behavior, but cannot be psychologically real. Ultimately, a sense of the unreality of ideal agent models can spur doubts about the very possibility of a cognitive science.

The other type of response to troubles with the idealizations is a *via media* strategy. After recognizing that nothing could count as an agent or person that satisfied *no* rationality constraints, one stops to wonder whether one must jump to a conclusion that the agent has to be ideally rational. Is rationality all or nothing, or is there some golden mean between unattainable, perfect unity of mind and utter, chaotic disintegration of personhood? The normative and empirical rationality models of Simon (1982) are among the earliest of this less stringent sort: the central principle is that, rather than optimizing or maximizing, the agent only “satisfices” its expected utility, choosing decisions that are good enough according to its belief-desire set, rather than perfect. Such modest coherence realistically is all that an agent ought to attempt, and all that can in general be expected. What amounts to a corresponding account for agent-constitutive rationality appears in Cherniak (1981), with a requirement of minimal, rather than ideal, charity on making sense of an agent’s actions. An even more latitudinarian conception can be found in Stich (1990). Related limited-resource models are now also employed in artificial intelligence (see BOUNDED RATIONALITY).

Moderate rationality conceptions leave room for the above-mentioned widely observed phenomena of suboptimal human reasoning, rather than excluding them as unintelligible behavior. We are, after all, only human. Indeed, these more psychologically realistic models can explain the departures from correctness as symptoms of our having to use more efficient but formally imperfect “quick but dirty” heuristic procedures. Formally correct and complete inference procedures are typically computationally complex, with surprisingly small-sized problem instances sometimes requiring vastly unfeasible time and memory resources. (To an extent, this practical intractability parallels, and extends, classical absolute unsolvability; see GÖDEL’S THEOREMS.) Antinomies like Russell’s paradox lurking at the core of our conceptual scheme can then be interpreted similarly as signs of our having to use heuristic procedures to avoid computational paralysis.

To conclude, some vigilance about unwarranted reification of cognitive architecture remains advisable. Just as attention has turned to evaluation of uncritical idealizing, scope continues for scrutiny of tacit assumptions in rationality models about psychologically realistic representational format (if any)—for example, the discussions reviewed above tend to presuppose agents as sentence-processors, rather than as, say, quasi-picture processors. Finally, the familiar uneasy coexistence of the intentional framework—having rationality at its core—with the scientific worldview is worth recalling. Yet probably much of the groundplan of our species' model of an agent is innate (see AUTISM and THEORY OF MIND); the framework therefore may be a ladder we cannot kick away. It is as if the scientific worldview can comfortably proceed neither with, nor without, an intentional-cognitive paradigm.

See also ANOMALOUS MONISM; COMPUTATIONAL COMPLEXITY; FOLK PSYCHOLOGY; INTENTIONAL STANCE

—Christopher Cherniak

References

- Cherniak, C. (1981). Minimal rationality. *Mind* 90: 161–183.
- Cherniak, C. (1986). *Minimal Rationality*. Cambridge, MA: MIT Press.
- Davidson, D. (1980). Psychology as philosophy. In D. Davidson, *Essays on Actions and Events*. New York: Oxford University Press.
- Dennett, D. (1978). Intentional systems. In *Brainstorms*. Cambridge, MA: MIT Press.
- Hempel, C. (1965). Aspects of scientific explanation. In *Aspects of Scientific Explanation*. New York: Free Press.
- Hintikka, J. (1962). *Knowledge and Belief*. Ithaca, NY: Cornell University Press.
- Quine, W. (1960). *Word and Object*. Cambridge, MA: MIT Press.
- Simon, H. (1982). *Models of Bounded Rationality*, vol. 2. Cambridge, MA: MIT Press.
- Stich, S. (1990). *The Fragmentation of Reason*. Cambridge, MA: MIT Press.
- Tversky, A., and D. Kahneman. (1974). Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124–1131.
- Von Neumann, J., and O. Morgenstern. (1944). *Theory of Games and Economic Behavior*. New York: Wiley

Rational Choice Theory

The theory of rational choice was developed within the discipline of economics by JOHN VON NEUMANN and Oskar Morgenstern (1947) and Leonard Savage (1954). Although its roots date back as far as Thomas Hobbes's denial that reason can fix our ends or desires (instrumental rationality), and David HUME's relegation of reason to the role of "slave of the passions," having no motivating force, via the utilitarians' definition of rationality as the maximization of "utility" and the neoclassical school of economics' theory of revealed preferences, rational choice theory (RCT) purports to be neutral relative to all forms of psychological assumptions or philosophies of mind. In this respect, its relevance for the cognitive sciences is problematic. However, its most recent developments have been marked by the discovery of paradoxes (Binmore 1987a, b; Campbell and Sowdon 1985;

Eells 1982; Gauthier 1988/89; Gibbard and Harper 1985; Kavka 1983; Lewis 1985; McClennen 1989; Nozick 1969; Rosenthal 1982) whose interpretation and resolution call for the return of the repressed: an explicit psychology of DECISION MAKING and a full-blown theory of mind. No wonder more and more cognitive scientists today (philosophers, artificial intelligence specialists, psychologists) participate, along with economists and game theorists, in the debates about RCT.

It is ironic that Savage's expected utility theory, in which most economists see the perfect embodiment of instrumental rationality, is a set of axioms, admittedly purely syntactic in nature, that constrain the rational agent's ends for the sake of consistency (see RATIONAL DECISION MAKING). For instance, her preferences must be transitive: if she prefers x to y and y to z , she *must* prefer x to z . If, no matter the state of the world, she prefers x to y , she *must* prefer x to y even in the ignorance of the state of the world (*sure-thing principle*). Savage proves that an agent whose preferences satisfy all the axioms of the theory chooses *as if* she were maximizing her expected utility while assigning subjective probabilities to the states of the world. It is not at all that her choices can be *explained* by her setting out to maximize her utility, because it is tautological, by construction, that the utility of x is larger to her than that of y if she chooses x over y . The claim is that agents whose preferences were not consistent (i.e., violated the axioms) could not achieve the maximal satisfaction of their ends.

This removal of all psychological content and motivational assumptions from the theory of utility is untenable. Consider the obvious possibility that preferences may change over time. Which of one's preferences should be subjected to the coherence constraints set by the theory? Only the occurrent ones, because future preferences are not motivationally efficacious now? Should we rather postulate second-order preferences that weigh future versus occurrent first-order preferences? Or are there (noninstrumental) *external* reasons that will do the weighing? Dispensing with a theory of mind proves impossible (Hampton 1998; Hollis and Sugden 1993).

According to RCT, an act is an assignment of consequences to states of the world, and the description of a consequence must include no reference to how that consequence was brought about. The only legitimate motivations are forward-looking reasons: only the future matters. Using an equipment just because one has invested a lot in it is taken to be irrational ("sunk cost fallacy"; see Nozick 1993). Experiments in cognitive psychology reveal that most of us commit that alleged fallacy most of the time, proving that we care about the consistency between past and present, maybe for the sake of personal identity (we violate as well Savage's axioms, especially the sure-thing principle; see Shafir and Tversky 1993; cf. also JUDGMENT HEURISTICS). Does that mean that we are irrational, or just that our mind works differently from what RCT, in spite of its proclaimed neutrality, presupposes?

When RCT is applied to a strategic setting, leading to GAME THEORY, some of its implications are plainly paradoxical. In an ideal world where all agents are rational, this fact being common knowledge (everyone knows it, knows