Mistakes and Rational Choice Theory Abstract

This paper concerns the concept of a mistake. Everyone makes mistakes, and each of us typically makes mistakes every day. Some of these mistakes are inconsequential; others are life-altering. We all do our best to learn from our mistakes and those of others so we do not repeat them in the future.

Yet, our best formal theories of choice do not have space for mistakes. According to rational choice theory, all behavior is interpreted as the agent's maximizing a utility function. We do not have the tools to identify which behaviors are the result of maximizing a utility function and which are not. In order to provide a plausible theory of human action, rational choice theorists need to make room in their theories for the notion of a mistake, but it is unclear how they could say which actions are mistakes and which ones are not.

I will not assume that mistakes are a theoretically unified kind. They are unified by what they are not: they are not normally functioning instances of our decision-making capabilities. There are many reasons why we make mistakes, and many kinds of mistakes. But what makes something a mistake is that it is an instance where someone does something that they prudentially ought not have done.

The paper will proceed in two parts. The first part produces an account of mistakes that is internal to RCT. In other words, I identify several kinds of explanations for mistakes that are consistent with the rational choice perspective on human behavior. Specifically, I identify four different kinds of mistakes: those owing to factual ignorance, those owing to failures of self-knowledge, those owing to preference change, and those owing to normative ignorance. Then, in the second section, I examine some more fundamental issues in philosophy of mind and philosophy of science. Here, I try to make sense of mistakes using resources external to RCT. I show how mistakes should be expected on a certain theory *of* RCT: Quinean holism. Attributing preferences to people is fundamentally an exercise in radical interpretation. We attribute mental states like preferences to people in roughly the same way that we attribute linguistic meanings to speakers. Organizing the messy and perhaps contradictory data of choice into a clean and orderly utility function is a perilous task, similar to the problem of the indeterminacy of translation and the underdetermination of theory by evidence. On this view, mistakes are episodes where not all of the desiderata for constructing a theory of a person's behavior are jointly satisfiable.