A Reason-Based Model of Rational Choice

(extended abstract)

Franz Dietrich and Christian List

May 22, 2009

The standard rational choice paradigm explains an individual's preferences by his beliefs and his fundamental desires. For instance, someone's preference for joining the army might be explained by certain beliefs about what life in the army is like and a desire for such a life. When the paradigm is spelled out formally, the objects of preferences (such as possible professions) are usually ranked according to their expected utility, derived using a probability function reflecting current beliefs and a utility function reflecting never-changing fundamental desires. In consequence, changes in preference can result only from changes in belief, not from any fundamental changes in desire or motivation.

This standard paradigm violates the intuitions of many and is frequently criticised. One shortcoming is that reasons and motivations play no explicit role. Some of the more fundamental preference changes that one can undergo seem to reach beyond information-learning and to involve a change in the reasons or goals by which one is fundamentally motivated. Such changes of motivating reasons may come in connection with a changing ability to abstractly represent certain aspects of the world (like the thirteenth move in a game) or to imagine certain qualitative aspects of the world (like feelings of complete loneliness). But standard rational choice models do not, or at least not explicitly, address these phenomena. Rather, as one can argue, they implicitly assume away limitations or changes in conceptualisation (by identifying the individual's representation of the world with the modeller's) or in imagination (by using invariant fundamental desires).

Criticisms of rational choice theory often suffer from not offering a formal alternative. The literature has of course modelled several bounded forms of

rationality; these are important in their own right, but they usually take the expected-utility paradigm as their starting point and introduce certain deviations from it, without introducing reasons or motivations into the model.

This paper proposes a formal reason-based model of preferences. The model explains an individual's preferences by the set of reasons that motivate him. The preference of our example individual for joining the army would be explained by the set of reasons that motivate him, such as service to his country, an athletic body, and comradeship. Preference *change* in our model thus stems not exclusively from new information but often also from a change of the set of motivating reasons. If our example individual suddenly loses his preference for joining the army and joins a charity, new reasons (such as worldwide justice) might have become motivating while others (such as an athletic body) might have lost their motivational power. Our notion of a '(motivating) reason' is essentially technical, and as such open to different interpretations and applications, like ones related to conceptualisation or imagination abilities.

We formulate two natural postulates on reason-based preferences, the first ensuring that preferences are determined by the motivating reasons and the second ensuring that preferences change in a coherent way as additional reasons become motivating. These two postulates are shown to imply a simple representation of individual preferences: a single binary relation (which ranks the consistent reason sets) is sufficient to generate all individual preferences across possible individual states (i.e., possible sets of motivating reasons).

Although our model of preferences looks very different from classical rational choice models – for instance, we do not generally reduce preference change to belief change – a valid question is whether our model simply redescribes classic informational preference change in different terms. More precisely, can a change in motivating reasons (as it occurs in our model) be reduced to a change in information in some suitably enriched model? Indeed, an orthodox critic might immediately suspect that our reason-based model of preference change describes a macroscopic phenomenon whose microscopic origin is an informational change; and he might attempt to formulate a classical rational choice model to which our model can be reduced (just as cooperative game theory can to some extent be reduced to non-cooperative game theory). Such a classical reduction of our model – if it is possible – would make our model less of a departure from orthodox rational choice theory (without making it redundant, since a macroscopic perspective has its own merits, such as parsimony-related merits). We briefly engage

with this important question, arguing against classical reducibility. We argue that our reason-based preference change model can instead be reduced to non-orthodox models of phenomena such as change of conceptualisation or imagination abilities. Our model is a macroscopic model that refrains from explaining *why* some reasons motivate and others don't; this is on purpose, in part because the model should not be committed to one of the potential microscopic explanations.

We do by no means intend to deny that many preference changes are information-driven, or to abolish the standard rational choice model. A comprehensive rational-choice theoretic model would explain someone's preferences by two factors, information and motivating reasons. While the present paper's model captures only the latter and standard rational choice models capture only the former, it is possible to merge both approaches to a comprehensive model of preference formation and change, as only briefly discussed at the end.