Where do preferences come from? A summary[*]

Franz Dietrich
CNRS & University of East Anglia

Christian List
London School of Economics

24 July 2013

1 Introduction

The paper to be presented at this conference (Dietrich and List 2013) sketches some basic ideas underlying a broader, ongoing decision-theoretic project. In that project, we aim to develop a new, general approach to decision theory which

(i) improves upon standard decision theory in both its idealized, 'rational' and its more psychologically informed, 'behavioural' variants (the former are associated with classical rational choice theory, the latter with behaviourial economics and economic psychology);

(ii) is widely applicable in the social sciences and in philosophy; and

(iii) provides a framework for expressing some key philosophical debates about the relationship between reasons and rational decisions, which are not adequately captured by standard formal decision theory.

One important task is the development of a theory of preference formation and preference change, since standard decision theory says very little about where an agent's preferences come from and when they might change. Here, we introduce a simple formal framework for modelling preference formation and preference change.[1]

2 Informal summary

The idea that a rational choice is (among other things) a choice based on reasons – perhaps subjective reasons – is a very natural one, but the notion of a reason is more or less absent from standard decision theory. In standard models, an agent has beliefs and preferences, formally modelled as subjective probabilities and utilities, and acts so as to satisfy his (or her) preferences according to his (or her) beliefs. While beliefs may be updated in light of new evidence, the agent's preferences – at least with respect to

[*] This summary is based on Dietrich and List (2013) and presents, in abridged and adjusted form, material from that paper, supplemented by some more general remarks about our ongoing research. We would kindly request that, in any references, the present summary be cited together with the original paper. Christian List wishes to acknowledge support from the Leverhulme Trust.

[1] Some important related contributions include work by logicians on the logic of preference and preference change. See, e.g., Hansson (2001), de Jongh and Liu (2009), Grüne-Yanoff and Hansson (2009), Liu (2010), and the historical contribution by von Wright (1963). More detailed references to related literature, including some related work in economics, are included in Dietrich and List (2003).

fully described outcomes – are typically assumed to be *fixed* and *exogenously given* (in Dietrich and List 2013, we cite a few exceptional works, many of them from outside mainstream economics). An agent's preferences are simply taken to be an essential but inexplicable feature of his personal identity. On this picture, preferences cannot be rationally assessed or criticized (provided they satisfy some minimal internal consistency constraints), and we cannot capture the idea that an agent's preferences may be the product of something more fundamental, such as the agent's reasons and his weighing of these reasons.

To overcome these limitations, we propose a 'property-based' account of preference formation. The central idea is that an agent's preferences are based on certain 'motivationally salient' properties of the alternatives over which the preferences are held. An agent's preferences may change as new properties of the alternatives become salient or previously salient properties cease to be salient. The motivationally salient properties serve as the reasons for the agent's preferences.

More precisely, the agent ranks different alternatives according to the way he 'weighs' the motivationally salient properties of these alternatives. This works as follows. For each alternative, the agent considers the set of motivationally salient properties of that alternative. An underlying 'weighing relation', defined as a binary relation over sets of properties, is then used to rank these property combinations relative to each other.

For example, when a consumer forms his preferences over different goods in a supermarket, such as different yoghurts, he could in principle consider a very large number of properties (characteristics) of these goods. In practice, however, he will only consider a small subset of these properties: the motivationally salient properties. This may include whether a yoghurt is cherry-flavoured, low-fat, and free from chemical additives, but exclude whether the yoghurt has an odd number of letters on its label (a totally irrelevant property) or whether it has been produced in an environmentally sustainable manner (something only an ethically oriented consumer will pay attention to). The consumer then determines his preferences over different yoghurts on the basis of his weighing relation over property combinations. He will most prefer the yoghurt with the 'highest-ranked' combination of motivationally salient properties. The consumer's preferences can change when new properties of the alternatives become motivationally salient, for example when he starts caring about environmental sustainability, or when previously salient properties cease to be salient, for example when he becomes less diet-conscious.

The theory also permits a normative reinterpretation. Under this reinterpretation, the focus is no longer on the properties the agent is *actually* motivated by, i.e., the *motivationally salient* properties, but instead on the properties the agent *ought ideally* to be motivated by, i.e., the *normatively relevant* properties. The appropriate weighing relation then captures, not the agent's *actual disposition* to weigh different property combinations relative to each other, but the way they *ought* to be weighed, according to some normative background theory.

3 The formal framework

In what follows, we give a brief exposition of the central formal framework and a simple axiomatic characterization result.

3.1 Preferences and properties

Let X be a non-empty set of fundamental objects of preference (e.g., fully described outcomes or consequences of actions, possible worlds, social states, bundles of goods, or policy platforms). The elements of X are mutually exclusive and jointly exhaustive. We call them *alternatives*.

We represent the agent's *preferences* by some order $\succsim$ on X (a complete and transitive binary relation), where $x \succsim y$ means 'the agent weakly prefers x to y'. As usual, $\succ$ and $\sim$ denote the strict and indifference parts of $\succsim$.

To address the question of how $\succsim$ is formed and when it may change, we introduce the idea that the agent's preferences depend on certain properties of the alternatives. Informally, a *property* is a characteristic that an alternative may or may not have. For example, being vegetarian is a property that a meal may or may not have. (For simplicity, we set aside non-binary properties; see, e.g., Dietrich and List 2011.) Formally, a *property* is an abstract object, P, which *picks out* a subset of X. (It need not be *identified* with that subset; two more distinct properties could in principle have the same extension of alternatives in X satisfying them.) Let $\mathcal{P}$ denote the set of all properties.

In forming his preferences, the agent focuses on some, but not necessarily all, properties of the alternatives. We call the properties that the agent focuses on the *motivationally salient* ones, and call the set of such properties, M, the agent's *motivational state*. Formally, $M \subseteq \mathcal{P}$. When a property is in M, this simply means that the agent pays attention to it; it does not imply that the property is satisfied by any particular alternative under consideration. Also, inclusion of a property in M does not mean that the agent is always positively, or always negatively, disposed towards alternatives with that property. It only means that whether or not an alternative has the property may sometimes make a difference to what the agent's preference in relation to that alternative is.

Motivational salience is a primitive notion of our framework. Which properties are motivationally salient for an agent in any context is a psychological question that our formalism alone cannot answer (here, empirical work is required). We write $\mathcal{M}$ to denote the set of all motivational states that are deemed psychologically possible for the agent. Formally, $\mathcal{M}$ is a non-empty set of sets of properties.[2]

To indicate notationally that the agent's preference order $\succsim$ depends on his motivational state M, we append the subscript M to the symbol $\succsim$. So, $\succsim_M$ denotes the agent's preference order in motivational state M. A full model of the agent requires the ascription of a *family* $(\succsim_M)_{M\in\mathcal{M}}$ of preference orders to the agent, consisting of one preference order $\succsim_M$for each motivational state $M \in \mathcal{M}$.

How exactly does $\succsim_M$ depend on M? The family of preference orders $(\succsim_M)_{M\in\mathcal{M}}$ is *property-based* if there exists a binary relation $\geq$ over property combinations (consistent sets of properties[3]) such that, for any motivational state $M \in \mathcal{M}$ and any alternatives

[2] By stating which specifications of M are included in $\mathcal{M}$, we can capture different assumptions about which properties can simultaneously become motivationally salient for the agent. This could include assumptions about 'crowding out' or 'crowding in' effects, whereby the motivational salience of some properties either rules out, or brings about, the motivational salience of others.

[3] A set of properties is *consistent* if there exists an alternative $x \in X$ which satisfies all of them.

$x, y \in X$,

$$x \succsim_M y \Leftrightarrow \{P \in M : x \text{ satisfies } P\} \geq \{P \in M : y \text{ satisfies } P\}.$$

When this definition applies, we say that x's having the properties in $\{P \in M : x$ satisfies $P\}$ and y's having the properties in $\{P \in M : y$ satisfies $P\}$ are the agent's *motivating reasons* for preferring x to y in state M. We call $\geq$ the agent's *weighing relation* over property combinations. The weighing relation ranks different property combinations relative to each other, indicating which property combinations – if salient – are 'preferable to' or 'better than' which others for the agent.

3.2 An example

A simple example illustrates our framework. Suppose an agent faces a choice between the following four alternatives:

S&H: a sweet and healthy cake, nS&H: a non-sweet and healthy cake,
S&nH: a sweet and unhealthy cake, nS&nH: a non-sweet and unhealthy cake.

For simplicity, suppose the only properties that may become motivationally salient are:

S: sweetness; H: healthiness.

Suppose further that any set of properties can in principle be motivationally salient, so that the set of all possible motivational states is

$$\mathcal{M} = \{\{\text{S,H}\}, \{\text{S}\}, \{\text{H}\}, \varnothing\}.$$

Now the agent's preferences across different $M \in \mathcal{M}$ might be as follows:

In state $M = \{\text{S,H}\}$: S&H $\succ_M$ nS&H $\succ_M$ S&nH $\succ_M$ nS&nH.
In state $M = \{\text{S}\}$: S&H $\sim_M$ S&nH $\succ_M$ nS&H $\sim_M$ nS&nH.
In state $M = \{\text{H}\}$: S&H $\sim_M$ nS&H $\succ_M$ S&nH $\sim_M$ nS&nH.
In state $M = \varnothing$: S&H $\sim_M$ nS&H $\sim_M$ S&nH $\sim_M$ nS&nH.

We must emphasize that this is just one example of what the agent's family of preference orders across different motivational states might be. (In general, the motivationally salient properties in the different states only *constrain* the agent's preferences in those states; they do not by themselves *determine* those preferences. The preferences are determined only together with the underlying weighing relation.)

The family of preference orders in the present example can be verified to be property-based, with respect to the following weighing relation:

$$\{\text{S,H}\} > \{\text{H}\} > \{\text{S}\} > \varnothing.$$

The example illustrates that, when the agent's preferences are property-based, a single weighing relation over property combinations suffices to induce the agent's entire family of preference orders across different motivational states.

3.3 An axiomatic characterization

We now offer an axiomatic characterization of property-based preferences. The following two axioms constrain the relationship between motivationally salient properties and preferences.

Axiom 1 *The agent is indifferent between any two alternatives whose motivationally salient properties are the same. Formally, for any two alternatives $x, y \in X$ and any motivational state $M \in \mathcal{M}$,*

$$\textit{if } \{P \in M : x \textit{ satisfies } P\} = \{P \in M : y \textit{ satisfies } P\}, \textit{ then } x \sim_M y.$$

Axiom 2 *If the agent's motivational state changes, in that additional properties become motivationally salient, the agent's preference between any alternatives satisfying none of the newly added properties remains unchanged. Formally, for any two alternatives $x, y \in X$ and any two motivational states $M, M' \in \mathcal{M}$ with $M' \supseteq M$,*

$$\textit{if neither } x \textit{ nor } y \textit{ satisfies any } P \in M' \backslash M, \textit{ then } x \succsim_M y \Leftrightarrow x \succsim_{M'} y.$$

In the main paper, we prove that, if the set of possible motivational states $\mathcal{M}$ satisfies a suitable closure condition, Axioms 1 and 2 characterize the class of property-based families of preference orders. Call $\mathcal{M}$ *intersection-closed* if, for all $M_1, M_2 \in \mathcal{M}$, we have $M_1 \cap M_2 \in \mathcal{M}$.

Theorem 1 *Suppose $\mathcal{M}$ is intersection-closed. Then the agent's family of preference orders $(\succsim_M)_{M \in \mathcal{M}}$ satisfies Axioms 1 and 2 if and only if it is property-based.*

Thus the two axioms guarantee that the agent's family of preference orders across motivational states can be represented by a single underlying weighing relation over property combinations. The present result is just one of several theorems than can be obtained in our framework.

3.4 The basic implication

According to our theory, the stable feature characterizing an agent is not the agent's preference order over the alternatives in X, but the agent's weighing relation over property combinations. On this picture:

- An agent *forms* his or her preferences by adopting a particular motivational state, i.e., by focusing – consciously or otherwise – on certain properties of the alternatives as the motivationally salient ones (and by adopting a weighing relation in the first place).
- An agent may *change* his or her preferences when the motivational state changes, i.e., when new properties of the alternatives become motivationally salient or others cease to be salient.

In the main paper, we consider in detail whether our theory is empirically testable and present some game-theoretic applications.

4 Concluding remarks

We conclude this summary with a few general remarks on the kinds of issues our theory is intended to shed light on.

4.1 Non-informational preference change

Standard decision theory has no difficulty explaining how an agent's preferences over some uncertain prospects (as opposed to fully described outcomes) can change as the agent changes his beliefs about the likelihood of various possible outcomes of those prospects. However, standard decision theory is unable to explain preference changes of a different kind: those driven by a change in (i) the agent's normative priorities, (ii) the salience of certain considerations, or (iii) the agent's conceptual scheme or mental representation of the decision alternatives. The proposed theory of property-based preference formation points towards a unified treatment of such phenomena.

4.2 Framing and nudging effects

Since Tvserky and Kahneman's seminal work (e.g., 1981), it is well known that people's choices are often influenced by subtle changes in how the decision options are described or framed. Standard decision theory struggles to explain such phenomena, which are sometimes interpreted simply (but incorrectly) as violations of rationality. The proposed framework allows us to go beyond this standard picture, by capturing the way in which different frames or descriptions activate different motivating reasons in an agent's preference-formation process.

4.3 Deliberation, beyond the exchange of information

While economists often think of deliberation just in terms of the exchange of information, philosophers, political scientists, and others hold that deliberation is much richer: it has many other aspects, from the consideration of arguments and the weighing of reasons to the hypothetical assumption of other agents' perspectives. The proposed theory provides a better language for capturing the content of deliberative processes. It is able to do so by (i) explicitly modelling the mechanisms by which reasons constrain preferences, (ii) permitting a formalization of the notions of motivating and normative reasons, and (iii) capturing the possibility that an agent's set of motivating reasons may change in response to normative reflection.

4.4 Ethical considerations in decision theory

While standard decision theory rests on a purely formal account of rationality that provides no resources for ethically assessing or criticizing an agent's preferences, philosophers and others are often interested in a more substantive account of rationality, under which we can assess an agent's motivations and distinguish between attitudes that are formally but not substantively rational, and attitudes that meet stronger substantive or moral constraints. The proposed theory is intended to capture such broader concerns and thereby to build a bridge between formal and substantive approaches to thinking about rationality.

5 References

de Jongh, D., Liu, F. (2009) Preference, priorities and belief. In T. Grüne-Yanoff and S. O. Hansson (eds.) *Preference Change: Approaches from Philosophy, Economics and Psychology.* Dordrecht (Springer): 85-108.

Dietrich, F., List, C. (2011) A model of non-informational preference change. *Journal of Theoretical Politics* 23(2): 145-164.

Dietrich, F., List, C. (2013) Where do preferences come from? *International Journal of Game Theory* 42(3): 613-637.

Grüne-Yanoff, T., Hansson, S. O. (2009) *Preference Change: Approaches from Philosophy, Economics and Psychology.* Dordrecht (Springer).

Hansson, S. O. (2001) Preference Logic. In D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*, 2nd ed., vol. 4. Dordrecht (Kluwer): 319-393.

Liu, F. (2010) Von Wright's '*The Logic of Preference*' revisited. *Synthese* 175(1): 69-88.

Tversky, A., Kahneman, D. (1981) The framing of decisions and the psychology of choice. *Science* 211(4481): 453-458.

von Wright, G. H. (1963) *The Logic of Preference.* Edinburgh (Edinburgh University Press).