A response to Baltag, Girard, and White

Franz Dietrich
CNRS & University of East Anglia

Christian List
London School of Economics

14 August 2013

1 Introduction

The aim of this short note is to offer some brief responses to the comments (on Dietrich and List 2013) that we have received from Alexandru Baltag and, in a coauthored piece, Patrick Girard and Shaun White. We would like to begin by thanking Baltag, Girard, and White for their thoughtful and generous comments. Given space constraints, unfortunately, we are not able to do full justice to all their interesting suggestions, thoughts, and questions here, but we hope to address at least some of their central points.

Although Baltag on the one hand and Girard and White on the other raise a number of distinct issues, there is some thematic overlap between them. Several of the comments concern the formal properties and substantive interpretation of the 'weighing relation', a central concept in our reason-based model of preference formation. In what follows, we first recapitulate the concepts of reason-based preferences and the underlying weighing relation and then address some of the questions Baltag, Girard, and White raise about those concepts and about our approach more generally.

2 Reason-based preferences and the weighing relation revisited

The aim of the formal framework developed in Dietrich and List (2013) is to model the relationship between an agent's preferences and his or her 'reasons' for holding those preferences. Preferences are represented by some (complete and transitive) order $\succsim$ on a set of alternatives X. Crucially, each alternative in X is conceptualized by the agent, not as a primitive object, but as a bundle of properties. A *property*, for our purposes, is a binary characteristic that an alternative may or may not have. At any given time, the agent is in a particular *motivational state*, defined by the set M of properties that the agent focuses on. Inclusion of a property in M only means that the agent cares about, or pays attention to, that property. It does not mean that he or she always likes, or always dislikes, the property; the property simply makes some difference to the agent's preferences in motivational state M. To indicate that the agent's preference order depends on M, we write $\succsim_M$ to denote the agent's preference order in motivational state M.

Assuming that $\mathcal{M}$ is the (non-empty) set of all motivational states deemed possible (e.g., compatible with the agent's psychology), we call the family $(\succsim_M)_{M\in\mathcal{M}}$ of preference orders across $M \in \mathcal{M}$ *property-based* if there exists an underlying binary relation $\geq$ over consistent sets (combinations) of properties such that, for any motivational state $M \in \mathcal{M}$ and any alternatives $x, y \in X$,

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

Property-basedness means that the agent's entire family of preference orders across different possible motivational states can be represented in terms of a single underlying binary relation $\geq$ over combinations of properties. We call this a *weighing relation*. It captures how good, or preferable, different combinations of properties are relative to each other, from the perspective of the agent. Any family of preference orders that satisfies two basic axioms (discussed in Dietrich and List 2013) is representable in this way.

3 Relation-theoretic properties of the weighing relation

Baltag notes that, in Dietrich and List (2013), we make two claims about the relation-theoretic properties of the weighing relation. Our first claim is that, for any property-based family of preference orders, the underlying weighing relation is *essentially unique*, meaning that it is unique on all pairs of property combinations that 'matter' for the agent's preferences, i.e., all pairs that co-occur in some set $X_M = \{\{P \in M : x \text{ satisfies } P\} : x \in X\}$ for $M \in \mathcal{M}$. Our second claim is that, even though the agent's preference order $\succsim_M$ in every motivational state $M \in \mathcal{M}$ is transitive, the underlying weighing relation can still be intransitive.

Baltag wonders whether these two claims are in tension. Specifically, he observes that the weighing relation restricted to each X_M is transitive, and suggests that our claim about the relation's possible intransitivity 'simply amounts ... to noticing that the weighing relation is not uniquely determined on the "irrelevant" sets (and hence can be taken to be non-complete or non-transitive on some of those sets).'

We agree with Baltag's first observation (that the relation $\geq$ restricted to each X_M separately is transitive), but disagree that the possible intransitivity is just due to the residual non-uniqueness. To explain this point, we recall an example from a different paper (Dietrich and List 2013a).

Consider a consumer choice over three alternatives (different cars):

a Monster Hummer, which is fast, big, but not environmentally friendly ($FB\neg E$);

a Sports Beetle, which is fast, not big, but environmentally friendly ($F\neg BE$);

a Family Hybrid, which is not fast, but big and environmentally friendly ($\neg FBE$).

Suppose that any of the three properties of a car, 'fast', 'big', and 'environmentally friendly', could serve as reasons for or against preferring it. Formally,

- property F (fastness) is satisfied by the Hummer ($FB\neg E$) and the Beetle ($F\neg BE$),
- property B (big) is satisfied by the Hummer ($FB\neg E$) and the Hybrid ($\neg FBE$),
- property E (environmentally friendly) is satisfied by the Beetle ($F\neg BE$) and the Hybrid ($\neg FBE$).

Suppose, further, that any subset of $\{F, B, E\}$ can potentially constitute a motivational state M. So $\mathcal{M}$ is the powerset of $\{F, B, E\}$. It can be checked that the following family of preference orders across $M \in \mathcal{M}$ is property-based (the left column shows the motivational state; the right column shows the corresponding preference order $\succsim_M$ in that state, with $\succ_M$ and $\sim_M$ denoting strict preference and indifference, respectively).

$$\begin{aligned}
M = \{F, B, E\} &: \text{Hummer} \sim_M \text{Beetle} \sim_M \text{Hybrid},\\
M = \{F, B\} &: \text{Hummer} \succ_M \text{Beetle} \succ_M \text{Hybrid},\\
M = \{B, E\} &: \text{Hybrid} \succ_M \text{Hummer} \succ_M \text{Beetle},\\
M = \{F, E\} &: \text{Beetle} \succ_M \text{Hybrid} \succ_M \text{Hummer},\\
M = \{F\} &: \text{Hummer} \sim_M \text{Beetle} \succ_M \text{Hybrid},\\
M = \{B\} &: \text{Hummer} \sim_M \text{Hybrid} \succ_M \text{Beetle},\\
M = \{E\} &: \text{Beetle} \sim_M \text{Hybrid} \succ_M \text{Hummer},\\
M = \varnothing &: \text{Hummer} \sim_M \text{Beetle} \sim_M \text{Hybrid}.
\end{aligned}$$

Importantly, the underlying weighing relation $\geq$ must have the following features (otherwise it would not generate the family of preference orders just shown):

$$\begin{gathered}
\{F, B\} \equiv \{B, E\} \equiv \{F, E\},\\
\{F, B\} > \{F\} > \{B\},\\
\{B, E\} > \{B\} > \{E\},\\
\{F, E\} > \{E\} > \{F\},\\
\{F\} > \varnothing,\\
\{B\} > \varnothing,\\
\{E\} > \varnothing.
\end{gathered}$$

Here $>$ and $\equiv$ denote the strict and indifference components of $\geq$. The intransitivity of $\geq$ is a direct consequence of the second, third, and fourth rows of the displayed list.

It is important to note that this intransitivity is not due to non-uniqueness. In the present example, the agent's family of preference orders is representable *only* by an intransitive weighing relation. This remains true even if the weighing relation is chosen in the 'sparsest' possible way, i.e., with the smallest number of relata with which the given family of preference orders can be generated. Formally, there may have to be cycles within the set $\bigcup_{M \in \mathcal{M}} (X_M \times X_M)$ of pairs that matter for the agent's preferences.

4 A further analysis of the weighing relation

We now move on to some of Girard and White's main questions about the weighing relation. First of all, they suggest developing an explicit logic to formalise the relationship between an agent's motivational states, his or her preferences in those states, and the underlying weighing relation. They also briefly describe some ingredients of such a logic. We agree that the development of a logic of this kind would be useful for a number of purposes, and, unfortunately, our paper does not offer one (instead, it offers a more 'ordinary', decision-theoretic framework in the tradition of the representation

theorems of von Neumann and Morgenstern, Savage, and others). We leave this issue as a challenge for future work, though we would like to draw attention to Osherson and Weinstein's related contributions (2012a,b).

We next turn to another important question raised by Girard and White: in which cases can the weighing relation, which is a binary relation over property *combinations*, be reduced to a binary relation over *individual properties*? Girard and White note one such case: the case in which the weighing relation has a *lexicographic structure* (though Girard and White do not use this term). In the lexicographic case, the weighing relation $\geq$ is generated by a *priority order*, R, over properties (a linear order). For notational simplicity, let $R(1), R(2), R(3)$ etc. denote the highest-ranked, second-highest-ranked, third-highest-ranked etc. properties with respect to R. (For simplicity, we assume that the total number of properties is finite.) For any two property combinations S_1 and S_2, we then define $S_1 \geq S_2$ if and only if either $S_1 = S_2$ or there is some n such that

- $R(n) \in S_1$ and $R(n) \notin S_2$, and
- for all $m < n$, $[R(m) \in S_1$ if and only if $R(m) \in S_2]$.

This yields a weighing relation $\geq$ induced by the priority order R. Interpretationally, properties here play the role of 'good-making features', and the priority order represents their order of importance. An approach along these lines is suitable for representing choices by checklists (e.g., Mandler, Manzini, and Mariotti 2012) or take-the-best heuristics (e.g., Gigerenzer et al. 2000).

In Dietrich and List (2013a), we consider another case in which the weighing relation can be represented in terms of further primitives: the *additive* case. Here we introduce a *weighing function* w over individual properties, which assigns to each individual property P a real number, $w(P)$, interpretable as the 'weight' of P. The weighing relation $\geq$ is now induced by the weighing function as follows. For any two property combinations S_1 and S_2, we define $S_1 \geq S_2$ if and only if

$$\sum_{P \in S_1} w(P) \geq \sum_{P \in S_2} w(P).$$

What do the lexicographic and additive cases have in common? The answer is relevant to Girard and White's question about when a weighing relation is reducible to either an ordering or a function *over individual properties*. Both the lexicographic and additive cases involve a *separable* weighing relation. Intuitively, in those cases, the 'valence' of a property for the agent – whether it counts in favour of or against an alternative when motivating – does not depend on which other properties are present. In the additive case, an even stronger condition of *additive separability* is met. Given space constraints, we set the formal details aside.

Separable weighing relations are certainly interesting and important, but it would be a loss of generality to assume that an agent's preference formation is always based on them. In this sense, Girard and White's comment draws attention to an important, but nonetheless special case.

5 The bigger picture

Finally, we would like to respond to some of Baltag's and Girard and White's comments about the 'bigger picture' underlying our approach. According to our framework, the stable feature of an agent is no longer the agent's preference order over the alternatives in X, as in standard rational choice theory, but the agent's weighing relation over property combinations. The variable feature is the agent's motivational state. Both Baltag and Girard and White raise some questions about this picture. These include the following:

(1) Is this picture sufficiently general? Does it need to be generalized further?
(2) Can the criticisms that are normally directed at the fixed-preference assumption of rational choice theory also be directed at our assumption of a fixed weighing relation?
(3) Are all changes between motivational states genuinely 'rational'?

Regarding question (1), there is indeed some scope for further generalization. The paper under discussion here (Dietrich and List 2013) allows the motivationally salient properties (those in M) to vary across different motivational states M, and thereby across different decision-making contexts that the agent might be in, but takes the properties themselves to be completely context-independent. This means that whether an alternative satisfies a given property does not depend at all on the context or situation in which the agent is confronted with that alternative. In ongoing work (Dietrich and List 2013b), we discuss the possibility that an agent's motivationally salient properties (those in M) may not just vary across different decision-making contexts, but that they may also include properties that refer to the context itself. To illustrate, consider Amartya Sen's famous example of a polite dinner party guest. This guest never chooses the largest piece of fruit offered to him or her, in order to avoid being greedy. So, at a superficial level, the agent seems to display different preferences over pieces of fruit in different situations. (Whether a particular apple is chosen – revealed-preferred in a given context – depends on which other pieces of fruit are also on offer.) However, the best explanation of what is going on here involves, not varying the motivational state M, but rather including the property of 'politeness' in M. Crucially, 'politeness' is a *relational* property: whether an alternative is 'politely choosable' depends not only on the alternative itself, but also on which other alternatives are available (a feature of the decision-making context). In Dietrich and List (2013b), we argue that these observations point towards two very different ways in which the decision-making context may make a difference: *context-variance* (here, the agent has different Ms in different contexts) and *context-regardingness* (here, some Ms may include properties that refer to the context, such as relational properties). It should be evident that introducing both kinds of context-dependence opens up a more general picture, of which the framework in Dietrich and List (2013) is a special case.

Regarding question (2), a critic is of course right to note that assuming a fixed weighing relation imposes a certain restriction. Methodologically, we believe, however, that a model of individual choice should not involve too many free variables; otherwise it would run the risk of becoming unfalsifiable. In addition, even when there are many free variables, each agent needs to be specified in terms of *some* fixed characteristics; otherwise it is unclear what explanatory role the ascription of agency plays (there still needs to be a sense in which we are dealing with a single agent). How restrictive or permissive the assumption of a fixed weighing relation is depends significantly on how

rich the set of properties is that we invoke for explanatory purposes. With sufficiently many properties – possibly allowing non-separable interaction effects between them – we may indeed be able to explain even complex shifts in the agent's evaluative dispositions. In Baltag's comment, he gives the example of someone who 'can fall in love, or fall out of love, just so, purely and simply, for no deeper reasons'. If, by 'reasons', we mean 'substantively rational and fully conscious reasons', we would of course agree; such is human psychology. However, if an agent's preferences change, we may still be able to attribute that change to some *cause*, not necessarily a substantively rational or conscious *reason*. We think that especially the more general version of our framework, in which contexts can influence an agent in two very different ways, can accommodate apparently non-rational preference changes, without having to suggest that they are 'uncaused'.

Finally, regarding question (3), we wish to note that the technical framework presented in Dietrich and List (2013) only allows us to assess the formal relationship between an agent's motivational state M and the corresponding preference order $\succsim_M$. (For example, does that relationship satisfy the two axioms characterizing property-basedness?) We do not wish to suggest that changes from one state M to another state M' are always substantively rational. Indeed, Girard and White provide a nice 'money pump' example of what can go wrong when an agent keeps vacillating between different motivational states. The more general framework in Dietrich and List (2013b) that we have briefly described can be used to draw a distinction between (i) those forms of context-dependence that can count as rational in some sophisticated sense (certain kinds of norm-following might fall into this category, as in Sen's politeness example) and (ii) those forms of context-dependence that are boundedly or 'sub-'rational. Girard and White's example would fall into the latter category.

We conclude by thanking Baltag, Girard, and White once again for their comments.

6 References

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

Dietrich, F., List, C. (2013a) A reason-based theory of rational choice. *Nous* 47(1): 104-134.

Dietrich, F., List, C. (2013b) Reason-based rationalization. Working paper, London School of Economics.

Gigerenzer, G., Todd, P. M., ABC Research Group (2000) *Simple Heuristics that Make Us Smart.* New York (Oxford University Press).

Mandler, M., Manzini, P., Mariotti, M. (2012) A million answers to twenty questions: Choosing by checklist. *Journal of Economic Theory* 147: 71-92.