An Axiomatic Approach to Revising Preferences (Extended Abstract)

Adrian Haret[1], **Johannes P. Wallner**[2]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]Institute of Software Technology, Graz University of Technology, Austria
a.haret@uva.nl, wallner@ist.tugraz.at

Abstract

Preferences play a central role in theories of decision making as part of the mechanism underlying rational choice: they show up in economic models of rational agency (Sen 2017), as well as in formal models of artificial agents expected to interact with the world and each other (Domshlak et al. 2011; Rossi, Venable, and Walsh 2011; Pigozzi, Tsoukiàs, and Viappiani 2016). Since such interactions take place in dynamic environments, it can be expected that preferences change in response to new developments.

In this paper we are interested in preference change occurring when new preference information becomes available and has to be taken at face value, thereby prompting a change in the prior preference. The change, we require, should preserve as much useful information from the prior preference as can be afforded. Preference change thus described is a pervasive phenomenon, arising in many contexts spanning the realms of both human and artificial agency. One prominent example is the distinguished tradition in Economics and Philosophy looking at examples of conflict between an agent's subjective preference (what we call here the prior preference π) and a second-order preference, often standing for a commitment or moral rule (what we call here the new preference information μ): subjective versus 'ethical' preferences (Harsanyi 1955), lack of will, or *akrasia* (Jeffrey 1974), moral commitments (Sen 1977), second-order volitions (Frankfurt 1988) and second-order preferences (Nozick 1994) all fall under this heading.

The same challenge can occur in technological applications, from updating CP-nets (Cadilhac et al. 2015) to changing the order in which search results are displayed on a page in response to user provided specifications, as well as, more generally, in issues related to the *alignment problem* (Russell 2019): an artificial agent dealing with humans will have to learn their preferences, but as it cannot do so instantaneously, it must presumably do so in intermediate steps, revising along the way. The following example illustrates the problem in its most basic form.

Example 1. *An online streaming service constructs a profile tailored to a particular user, according to which the arthouse movie (a) is preferred to the biopic (b), which is preferred to the comedy (c), and thus displays them in this order, encoded here with the preference statement $\pi = (a \succ b) \wedge (b \succ c)$. When the user volunteers information to the effect that they find the comedy better than the arthouse movie, i.e., new information $\mu = (c \succ a)$, the streaming service must revise its model of the user's preference: it has to place c before a and, in order to display the alternatives in a neat linear fashion, it must decide on a position to slot b into. Preferences π and μ, together with possible values for the revised result, e.g., $\pi_1 = (c \succ a) \wedge (a \succ b)$, $\pi_2 = (c \succ a) \wedge (b \succ c)$ and $\pi_3 = (c \succ a) \wedge (c \succ b) \wedge (b \succ a)$, are depicted in Figure 1. Intuitively, π_3 veers far (too far) away from the input preference π, in that it does not keep any of the still permissible comparisons contained in π, and should arguably be excluded, while π_1 and π_2 are viable contenders. If we go further and insist on a decision between π_1 and π_2, we can take stock of the information relayed by either choice. Accepting π_1 involves giving up the comparison of b over c, and we may surmise this is because the comparison of a over b is given up more reluctantly: preference of the arthouse movie over the biopic is more intense! Acceptance of π_2 implies the opposite: b over c is now the stronger preference. Thus, restricting the output of revision to a single linear order suggests that the choice can be rationalized using an implicit preference order over the comparisons.*

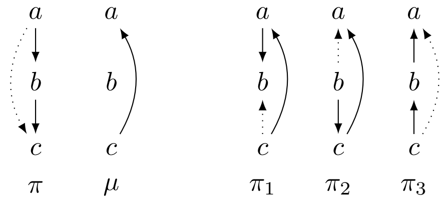

Figure 1: Revising π by μ can be thought of a choice between which comparisons to keep and which to give up.

Thus, whether it is the internal conflict of a moral agent or a content provider aiming for a better user experience, many cases of preference change involve a conflict between two types of preferences, one of which has priority. But, despite the fact that the problem is often signaled, a principled approach to how to handle it is often overlooked.

Our aim in the paper is to formalize the type of reasoning illustrated in Example 1 by rationalizing preference change as a type of choice function that utilizes the information provided by the prior preference in adapting itself to new information. In particular, we combine techniques from standard belief change with Sen's insight that conflicts among preferences should be resolved using preferences over the preferences themselves (Sen 1977), and put forward two models for preference revision: *irresolute revision* and *resolute revision*.

Irresolute Preference Revision An *irresolute preference revision operator* $\circ$ is a function taking as input two preference statements, typically denoted π and μ, and standing for the agent's prior and newly acquired preference information, respectively, and returning a *set* of preference statements, denoted $\pi \circ \mu$. The representation of the result as a set of formulas is a slight departure from established revision practice, but has precedent in belief change applied to formalisms other than propositional logic, e.g., in work on the aggregation of abstract Argumentation Frameworks (Delobelle et al. 2016). Intuitively, $\pi \circ \mu$ can be interpreted as a range of options, all of which, together, represent the agent's adjusted preferences in light of new information μ. A model of a preference statement π is a linear order ℓ that satisfies every atomic preference statement in π.

An important device for generating concrete revision operators is a *distance* d between linear orders. A typical distance we will use here is the *Kendall tau distance* d_τ (Kendall and Gibbons 1990) defined as $d_\tau(\ell_1, \ell_2) = |\{xy \in \ell_1 \mid yx \in \ell_2\}|$, i.e., as the number of disagreements (inverted pairs of alternatives) between ℓ_1 and ℓ_2. Less discriminating, the *drastic distance* d_D is defined as $d_D(\ell_1, \ell_2) = 0$, if $\ell_1 = \ell_2$, and $k > 0$, otherwise.

A distance-based revision operator $\circ^d$ works by selecting the models of μ that are overall closest to the models of π according to d. Notably, the allowing the result to be a set of preference statements allows us to represent operators that would otherwise not fit into the framework.

Example 2. *For $A = \{a, b, c\}$ and $\pi = (a \succ b) \wedge (b \succ c)$, $\mu = (c \succ a)$, we have that $[\pi] = \{abc\}$ and $[\mu] = \{cab, cba, bca\}$. The Kendall tau distances between the model of π and the models of μ are $d_\tau(abc, cab) = 2$, $d_\tau(abc, cba) = 3$, $d_\tau(abc, bca) = 2$, and thus $[\pi \circ^\tau \mu] = \{cab, bca\}$. We can represent $[\pi \circ^\tau \mu]$ using preference statements $\pi_1 = (c \succ a) \wedge (a \succ b)$ and $\pi_2 = (b \succ c) \wedge (c \succ a)$, noting that $[\pi_1] \cup [\pi_2] = \{cab\} \cup \{bca\} = \{cab, bca\}$, with $\pi \circ^\tau \mu = \{\pi_1, \pi_2\}$.*

At the same time, there is no preference formula π' such that $[\pi \circ^\tau \mu] = [\pi']$.

Our main result in this section is a representation theorem showing that irresolute revision operators satisfying a set of AGM-like postulates can be represented using familiar faithful assignments on the semantic side. We also show that any distance-based operator where the distance satisfies a set of intuitive properties (a class that includes the Kendall-tau distance, but not the drastic distance) *cannot* be represented with a single preference statement as output.

Resolute Preference Revision In the wake of the aforementioned results we want to understand how to think about preference revision operators where the formats of the prior and revised preference information are the same. Thus, a *resolute preference revision operator* $\circ$ is a function taking as input a complete (i.e., with a single preference statement as model) and a standard preference statement, typically denoted by λ and μ, respectively, and returning a complete preference statement, denoted by $\lambda \circ \mu$.

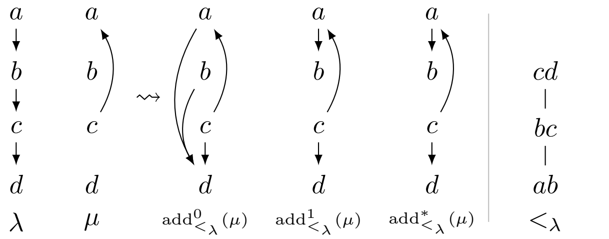

Figure 2: Revision according to a faithful linear order $<_\lambda$. Lower comparisons are better. Comparisons inferred by transitivity are omitted.

We define resolute preference revision operators starting with a linear order $<_\lambda$ on the adjacent comparisons of λ, and iteratively adding as many of these comparisons to μ as possible.

Example 3. *Take $\lambda = (a \succ b) \wedge (b \succ c) \wedge (c \succ d)$, $\mu = (c \succ a)$, with $[\lambda] = \{abcd\}$, $[\mu] = \{cabd, cbad, bcad, cadb, \dots\}$, with $<_\lambda$ depicted in Figure 2. The order $\lambda \circ \mu$ is assembled in steps, starting with the comparisons in μ and the cyclic-free part of λ, $(\{ca\} \cup \{ad, bd, cd\})^+$, depicted in Figure 2. Then, in the first step ab is added; the second step tries to add bc, but finds that this leads to a cycle with the previously added comparisons; the third step adds cd, which had been added already, and the process stops. The result is $[\lambda \circ^f \mu] = \{cabd\}$.*

The main result in this section is a representation theorem showing that this process can be axiomatized using a set of AGM-like postulates and rankings on the adjacent comparisons on λ.

References

Cadilhac, A.; Asher, N.; Lascarides, A.; and Benamara, F. 2015. Preference change. *Journal of Logic, Language and Information* 24(3):267–288.

Delobelle, J.; Haret, A.; Konieczny, S.; Mailly, J.; Rossit, J.; and Woltran, S. 2016. Merging of Abstract Argumentation Frameworks. In *Proceedings KR 2016*, 33–42.

Domshlak, C.; Hüllermeier, E.; Kaci, S.; and Prade, H. 2011. Preferences in AI: An Overview. *Artificial Intelligence* 175(7-8):1037–1052.

Frankfurt, H. G. 1988. Freedom of the Will and the Concept of a Person. In *What is a person?* Springer. 127–144.

Harsanyi, J. C. 1955. Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy* 63(4):309–321.

Jeffrey, R. C. 1974. Preference among preferences. *Journal of Philosophy* 71(13):377–391.

Kendall, M., and Gibbons, J. D. 1990. *Rank Correlation Methods*. New York: Oxford University Press.

Nozick, R. 1994. *The Nature of Rationality*. Princeton University Press.

Pigozzi, G.; Tsoukiàs, A.; and Viappiani, P. 2016. Preferences in artificial intelligence. *Annals of Mathematics and Artificial Intelligence* 77(3-4):361–401.

Rossi, F.; Venable, K. B.; and Walsh, T. 2011. *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Sen, A. K. 1977. Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs* 317–344.

Sen, A. K. 2017. *Collective Choice and Social Welfare: Expanded Edition*. Penguin UK.