Salient Preference Dynamics: A Model of Attention-Driven Preference Change

David Hyland
University of Oxford
david.hyland@cs.ox.ac.uk

Sarit Kraus
Bar-Ilan University and
University of Oxford
sarit@cs.biu.ac.il

Michael Wooldridge
University of Oxford
mjw@cs.ox.ac.uk

Abstract

A core challenge for real-world agents is determining which aspects of a decision problem are most relevant – that is, deciding what to pay attention to. Salience profoundly influences human decision-making, often leading to choices that diverge from normative rationality. In this extended abstract, we sketch a formal model that captures how salience modulates preferences and drives changes in preference. We provide a representation theorem characterising preference relations shaped by salience-weighted features. This framework provides a mathematical starting point for studying preference change under attentional modulation and for designing systems that can effectively guide human decision-making in complex, multi-faceted environments. At the same time, we highlight important considerations about the responsible design of influence strategies, particularly in contexts where subtle preference shaping may have unintended or harmful consequences.

Understanding and persuading other agents is a fundamental challenge that social beings face. To address this challenge, it is helpful to understand how an agent's preferences may change. As a starting point, Jeffrey [4] identifies two significant sources of preference change – valuational and doxastic. *Valuational preference change* refers to a change in the underlying utilities that an agent associates with different outcomes, whereas *doxastic preference change* results from a change in the agent's beliefs about (the likelihood of) different outcomes [6]. Valuational preference change may occur, e.g., by acquiring a taste for certain kinds of food through repeated exposure, or enjoying a sport more as one becomes more proficient at it. Examples of doxastic preference change include not wanting to purchase a particular product after learning that it was unethically sourced, or not wanting to visit the beach after learning that it is raining. Standard models of preference change often assume that agents possess fully resolved beliefs and preferences. However, bounded agents routinely fail to consider certain variables entirely – a phenomenon formally captured in epistemic logic as unawareness [2]. Building on this recognition of cognitive limits, we explore an alternative source of preference change: that of *salience weighting*. We propose a model for representing such weightings by modelling preferences over alternatives as a sum of individual utility functions for each feature weighted by its "salience". We represent salience as a probability distribution over feature weights, which, while distinct from lotteries over outcomes, allows us to apply tools from probability theory and belief updating to model the evolution of preferences under attentional shift.

Researchers in several disciplines, ranging from neuroscience to economics, have appreciated the influence of salience on preferences. For example, Schonberg and Katz [8] propose that attention-related regions of the brain contribute to non-reinforced preference change, i.e., changes in preference in the absence of external reinforcement. In his Nobel lecture, Daniel Kahneman asserted that

"the effects of salience and anchoring play a central role in treatments of judgment and choice" [5]. Salience typically refers to the quality of being noticeable, significant, or prominent to an agent. It relates to how agents prioritise different information or features relative to each other, and therefore takes into account not just *individual* features of an agent's sensorium, but a whole *collection* of them. When making real-world decisions, agents often face multiple competing factors that must be balanced against each other. This raises a key question: how do decision-makers prioritise different aspects of their options? For example, in a home-buying scenario, the prospective customer faces a richly multifaceted decision-making problem, with features like price, size, facilities, neighbourhood, etc., all critically factoring into their preferences [9]. In this context, a real estate agent typically interacts with the buyer, pointing out different features of different homes. Depending on their incentives, the real estate agent may have their own preferences over which property is finally selected, and they may try to guide the decision-maker towards choosing these properties. Another ubiquitous instance of this is the formation of voting preferences for political parties/candidates, who are primarily characterised by where they stand on different issues, including taxation, foreign policy, and healthcare. Different political candidates aim to shape the public's perception of themselves by drawing the attention of their target audience to what are perceived to be their strengths on particular issues and the apparent weaknesses of other candidates, thereby elevating their relative status [10]. What is common to these scenarios is that 1) agents making a choice must find some way to combine preferences about each feature into an *overall* preference among alternatives, and 2) agents' preferences can be swayed even in the absence of new information or a change to their underlying values. We aim to develop a simple model that can capture these occurrences of preference change through the lens of salience.

As a third example that we will return to throughout this text, consider a decision many students face when enrolling in university – selecting their degree. Such a task is highly nontrivial, has significant implications for the student's later trajectory in life, and contains many different features that must be weighed against each other [3]. Among other things, factors contributing to the decision of the prospective student include the median salary of graduates from each discipline, the difficulty of the course, the course fees, the typical length of time to completion, etc. Naturally, the student may have preferences over each of these features, but when it comes to making a decision, they must combine these preferences into an overall ranking of the alternatives.

1 A Model of Salient Preferences

Here, we develop a basic model that captures the influence of salience on preferences in the context of a multi-attribute decision making (MADM) problem [12, 13].

Feature Space and Alternatives. Consider an agent facing a choice among a set of *alternatives*, denoted $X = X_1 \times \cdots \times X_m$ is a product space of m *feature spaces*, where each $X_j \subseteq \mathbb{R}$ is an interval. Each alternative $x = (x_1, \ldots, x_m)$ is characterised by a unique combination of *feature instances* $x_j \in X_j, j \in \{1, \ldots, m\}$.

Salience. Salience is modelled as the relative importance of different features to the current preferences of the agent. Here, we assume that salience operates at the level of features, as opposed to value functions, which apply to feature instances. In particular, we let $s = (s_1, \ldots, s_m)$ be a *salience distribution*, which is a categorical distribution over the m features, and we write $S = \Delta^{m-1} = \left\{ s \in \mathbb{R}_+^m : \sum_{j=1}^m s_j = 1 \right\}$ for the set of all salience distributions.

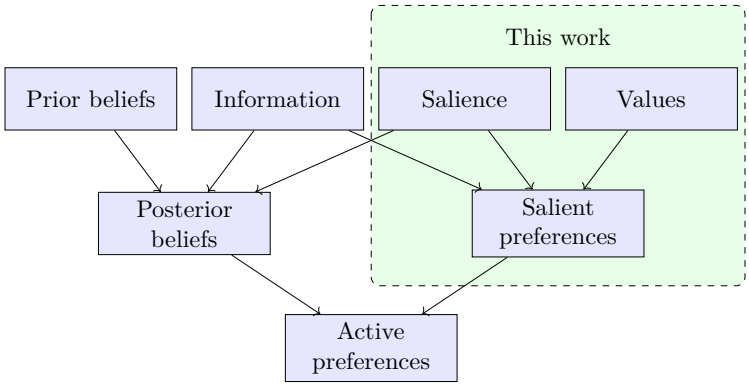

Figure 1: Basic schematic of the factors contributing to an agent's active preferences. As a first approximation, this can be split into a combination of beliefs, subjective values, and mechanisms including information and salience that interact with the two to form the agent's active preferences. Prior beliefs refer to the agent's beliefs coming into a particular decision-making context. Information refers to new data that the agent receives from its external environment or other agents, and internally generated information, e.g., outputs of deliberation processes. Salience represents the weighting placed on different features of the alternatives under consideration (irrespective of their actual instantiations), which modulates the agent's immediate beliefs and preferences. Values refer to the underlying preferences an agent possesses concerning different instances of each feature separately. From these basic building blocks, posterior/derived beliefs are formed by updating prior beliefs with any previously unincorporated information, which are then modified by saliences. Similarly, salient preferences result from information and salience interacting with underlying values. Finally, these are combined to form active preferences, which guide the agent's decisions.

1.1 Salient Preference Relations and Their Representation

To model the effect of salience on preferences, we augment the standard multi-attribute utility theory by introducing additional axioms. In particular, considering the set of all possible salience distributions, we assume the existence of a family of preference relations $\{\succeq_s : s \in S\}$, indexed by these distributions. We call this object a set of *salient preference relations* over X. Let δ_j be the salience distribution that allocates a salience probability of 1 to feature X_j. We write $x \sim_s y$ iff $x \succeq_s y$ and $y \succeq_s x$, and $x \succ_s y$ iff $x \succeq_s y$ and not $y \succeq_s x$.

We are now ready to introduce a set of axioms that establish the mathematical foundations for our model. These axioms provide necessary and sufficient conditions for representing these preferences through salience-weighted utility functions. A set of salient preference relations satisfies all these axioms iff it can be represented as a weighted sum of feature-specific utilities, where the weights correspond to salience values.[1]

1. *Completeness*: For any two alternatives $x, y \in X$ and any salience distribution $s \in S$, it holds that $x \succeq_s y$ or $y \succeq_s x$.
2. *Transitivity*: For any salience distribution $s \in S$, if $x \succeq_s y$ and $y \succeq_s z$, then $x \succeq_s z$.
3. *Alternative Continuity*: For any salience distribution $s \in S$ and any alternative $x \in X$, the sets

$$\{y \in X : x \succeq_s y\} \qquad \text{and} \qquad \{y \in X : y \succeq_s x\}$$

[1]We assume for simplicity that $m \geq 3$, with the special case of $m = 2$ omitted for brevity.

are closed in X.

4. *Pure-Salience Isolation*: For each $i \in \{1, \ldots, m\}$ and all $x, y \in X$,

$$x_i = y_i \Rightarrow x \sim_{\delta_i} y.$$

5. *Affine Salience Dependence*: Fix a reference alternative $x^0 \in X$. For each $s \in S$, let U_s be a continuous representation of $\succeq_s$ (whose existence is guaranteed by Axioms 1–3 [1]), which is normalised by setting $U_s(x^0) = 0$. Assume that these normalised representations can be chosen so that for every $x, y \in X$, the function

$$D_{x,y}(s) := U_s(x) - U_s(y),$$

is affine on S, i.e., for all $s, t \in S$ and all $\lambda \in [0, 1]$, we have

$$D_{x,y}(\lambda s + (1 - \lambda)t) = \lambda D_{x,y}(s) + (1 - \lambda) D_{x,y}(t).$$

6. *Nontriviality*: For each feature j, there exist $x_j, y_j \in X_j$ with $x_j \neq y_j$, some $x_{-j} \in X_{-j}$, and some $s \in S$ with $s_j > 0$ such that

$$(x_j, x_{-j}) \succ_s (y_j, x_{-j}).$$

Theorem 1. *For a family $\{\succeq_s : s \in S\}$, the following are equivalent:*

(i) Axioms 1–6 *hold.*

(ii) There exist continuous non-constant functions

$$v_j : X_j \to \mathbb{R} \qquad (j = 1, \ldots, m)$$

such that for all $x, y \in X$ and all $s \in S$,

$$x \succeq_s y \iff \sum_{j=1}^{m} s_j v_j(x_j) \geq \sum_{j=1}^{m} s_j v_j(y_j).$$

The theorem captures the desired separation between stable feature-level values and unstable attentional weights. The functions v_j do not vary with s; only the aggregation weights do. Consequently, a preference reversal between x and y can occur because the salience distribution changes, without any change in the values $v_j(x_j)$ or $v_j(y_j)$.

Example 1. *Returning to the enrollment example with the formal model articulated, suppose for simplicity that the student's alternative space is characterised by two dimensions - expected salary and average difficulty. Now consider two alternatives available to the student – computer science (CS) and medicine, which we will label x^1 and x^2, respectively. Again, simplifying considerably, suppose that the feature instances are given in Table 1. Under a uniform distribution, i.e., one that assigns equal weighting to all features ($s_1 = (0.5, 0.5)$), we see that medicine ranks slightly higher than CS with a total score of* 45*, compared to* 35*. However, under a salience distribution that places higher weight on the time or difficulty features, e.g., $s_2 = (0.2, 0.8)$, we see that the salience-weighted utility of CS comes to* 26*, whereas medicine receives a score of only* 24*. This illustrates how preferences may reverse under different salience distributions, where we have $x^2 \succeq_{s_1} x^1$ and $x^1 \succeq_{s_2} x^2$.*

Feature i	CS (x^1)	Med (x^2)	$v_i(x_i^1)$	$v_i(x_i^2)$
Salary ($)	100k	140k	50	80
Difficulty (/5)	4	5	20	10

Table 1: Feature instances for two alternatives in Example 1 and their associated values. In this example, the medicine degree offers a higher salary than the CS degree, but is also considered more difficult to complete. Depending on which features the agent pays attention to, either option may be seen as more attractive than the other.

2 Discussion

In summary, this extended abstract has introduced a formal framework for modelling how salience shapes preferences and drives preference change that does not rely on belief updating or value modification, but rather on shifts in attention allocation. Our key contribution is a representation theorem for preference relations indexed by salience distributions, demonstrating how salience-weighted features can combine to form overall preferences under a single set of value functions.

From the perspective of applications, this work provides foundations for designing AI systems that can effectively learn about and shape human preferences in support of individual and societal objectives. For example, Tiefenbeck et al. [11] discuss how making the impact of one's decisions more salient in real time, in the context of resource consumption, can promote more conservative resource-use behaviours, leading to a 22% reduction in energy consumption for targeted behaviours. In addition, Schenk [7] argues that strategically making use of *low* saliences can be effective in achieving desired outcomes, e.g., in the process of raising government revenue through taxes. However, these applications also highlight important considerations around their responsible deployment.

The ability to systematically influence preferences through the shaping of attention, while potentially beneficial when aligned with human values and intentions, could also enable concerning forms of unwanted manipulation if misused. The ability to influence preferences through attention modulation may be particularly concerning because these effects can occur without changing underlying values or beliefs. This suggests the need for more sophisticated governance frameworks for preference-shaping AI systems that incorporate insights from mathematical and computational models. If attention does play a significant role in human preferences, as the literature has pointed to from many directions, we need more precise technical definitions of concepts like preference stability to develop principled constraints on allowable influence strategies and to develop practices to defend ourselves against unwanted influence. Additionally, the possibility of unintended consequences when modifying attention patterns suggests that it would be prudent to develop robust validation protocols before deploying such systems.

This extended abstract takes steps toward a formal understanding of attention-driven preference change and suggests important directions for future work. As AI systems become increasingly capable of learning about and influencing human preferences, frameworks like the one presented here may be useful for ensuring that they do so in ways that enhance rather than diminish human agency. By better understanding how attention shapes preferences, we can build systems that help people align their immediate choices with their deeper values and aspirations. This research direction opens up possibilities for AI systems that could help humans navigate complex situations more effectively while promoting autonomy. The challenge ahead lies in thoughtfully developing these capabilities to empower humans in ways that align with their authentic preferences and well-being.

References

[1] Debreu, G. *Representation of a preference ordering by a numerical function.* Yale University, 1953.

[2] Fagin, R., and Halpern, J. Y. Belief, awareness, and limited reasoning. *Artificial intelligence 34*, 1 (1987), 39–76.

[3] Gillis, A., and Ryberg, R. Is choosing a major choosing a career or interesting courses? an investigation into college students' orientations for college majors and their stability. *Journal of Postsecondary Student Success 1*, 2 (2021), 46–71.

[4] Jeffrey, R. C. A note on the kinematics of preference. *Erkenntnis* (1977), 135–141.

[5] Kahneman, D. Maps of bounded rationality: A perspective on intuitive judgement and choice.

[6] Lang, J., and Van Der Torre, L. Preference change triggered by belief change: A principled approach. In *International Conference on Logic and the Foundations of Game and Decision Theory* (2008), Springer, pp. 86–111.

[7] Schenk, D. H. Exploiting the salience bias in designing taxes. *Yale J. on Reg. 28* (2011), 253.

[8] Schonberg, T., and Katz, L. N. A neural pathway for nonreinforced preference change. *Trends in Cognitive Sciences 24*, 7 (2020), 504–514.

[9] Selten, R. Aspiration adaptation theory. *Journal of mathematical psychology 42*, 2-3 (1998), 191–214.

[10] Shaki, J., Aumann, Y., and Kraus, S. Voter priming campaigns: Strategies, equilibria, and algorithms. *AAAI* (2025).

[11] Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., and Staake, T. Overcoming salience bias: How real-time feedback fosters resource conservation. *Management science 64*, 3 (2018), 1458–1476.

[12] Von Winterfeldt, D., and Fischer, G. W. Multi-attribute utility theory: models and assessment procedures. In *Utility, Probability, and Human Decision Making: Selected Proceedings of an Interdisciplinary Research Conference, Rome, 3–6 September, 1973* (1975), Springer, pp. 47–85.

[13] Zanakis, S. H., Solomon, A., Wishart, N., and Dublish, S. Multi-attribute decision making: A simulation comparison of select methods. *European journal of operational research 107*, 3 (1998), 507–529.