Beyond Theory and Data in Preference Modeling: Bringing Humans into the Loop

Thomas E. Allen[1], Muye Chen[2], Judy Goldsmith[1], Nicholas Mattei[3], Anna Popova[4], Michel Regenwetter[2], Francesca Rossi[5], and Christopher Zwilling[1]

[1] University of Kentucky, Lexington, KY, USA
{thomas.allen, goldsmit}@cs.uky.edu
[2] University of Illinois, Urbana–Champaign, IL, USA
{mchen67, regenwet, zwillin1}@illinois.edu
[3] NICTA and UNSW, Sydney, Australia
nicholas.mattei@nicta.com.au
[4] Dell Research Labs, Austin, TX, USA
anna_popova@dell.com
[5] University of Padova, Padova, Italy
frossi@math.unipd.it

Abstract. Many mathematical frameworks aim at modeling human preferences, employing a number of methods including utility functions, qualitative preference statements, constraint optimization, and logic formalisms. The choice of one model over another is usually based on the assumption that it can accurately describe the preferences of humans or other subjects/processes in the considered setting and is computationally tractable. Verification of these preference models often leverages some form of real life or domain specific data; demonstrating the models can predict the series of choices observed in the past. We argue that this is not enough: to evaluate a preference model, humans must be brought into the loop. Human experiments in controlled environments are needed to avoid common pitfalls associated with exclusively using prior data including introducing bias in the attempt to clean the data, mistaking correlation for causality, or testing data in a context that is different from the one where the data were produced. Human experiments need to be done carefully and we advocate a multi-disciplinary research environment that includes experimental psychologists and AI researchers. We argue that experiments should be used to validate models. We detail the design of an experiment in order to highlight some of the significant computational, conceptual, ethical, mathematical, psychological, and statistical hurdles to testing whether decision makers' preferences are consistent with a particular mathematical model of preferences.

1 Introduction

In the AI world of preference modeling, researchers often test their preference framework, particularly in the realm of recommendation systems and other decision support systems. However, most of the testing focuses on usability and

functionality. Almost none that we are aware of looks at whether humans actually act the way a certain preference model states; i.e., test the underlying assumptions of the model itself. Interest in testing preference models proposed in computer science began, for us, when thinking about conditional preference networks (CP-nets) [6]. Although there are many hundreds of papers on CP-nets, none that we know of has looked at actually eliciting CP-nets from non-computer scientists, nor done choice-based tests to see if people act in a manner consistent with having an underlying CP-net preference structure. In this paper we describe both the process and the challenges that go into designing and implementing a human subjects experiment to test, for instance, the validity of CP-nets. We argue that human subjects experiments are an important opportunity for both interdisciplinary collaboration as well as extending the scope and impact of preference research in computer science.

Even within the work on preference elicitation, we have noticed a focus on optimization (see, e.g., [7]) to make the process fast and not too invasive for the user. While we celebrate the increasing libraries of preference data available, such as PrefLib [44], we also have concerns about the efficacy of using those data alone for validating preference models. In particular, we see many models validated on the Sushi Dataset [30], e.g. [25], which was generated for a very particular scenario and yet is now exploited for tests in fundamentally different settings. When we generalize or attempt to switch the domain of some data we introduce bias, which can potentially lead to spurious conclusions about the methods under study [53]. There are usability studies for preference elicitation software (e.g., [9,52]) and humans are being brought into the loop in recommender systems (e.g., [28, 66]). These studies are crucial steps and the efforts should be rewarded and expanded within the broader communities that work with preferences. Running good tests with human subjects is necessary and nontrivial.

When we say that we advocate for studies with human subjects, by this we do not mean tests involving introspection. There is an urban legend in AI that the early work on chess involved asking chess players to introspect, and that this destroyed their intuitive processes. This likely refers to De Groots' work on chess:

> *"The only way of working with 'systematic introspection' would have been to interrupt the process after, say, every two minutes in order to have the subject introspect, and then continue. A few preliminary trials, however, with the author as subject showed this technique to be relatively ineffective as well as extraordinarily troublesome. After each interruption one feels disturbed and cannot continue normally. Apart from being unpleasant for the subject the technique is highly artificial in that it disrupts the unity of the thought process* [16, pp. 80–81]."

If we were to ask athletes to pay active attention to every body movement during peak performance, quite plausibly they would either disregard our instructions or fall short of peak performance due to a lack of focus. This is why athletes have coaches who monitor them. Likewise, asking decision makers to divert attention and memory resources away from their task in order to monitor

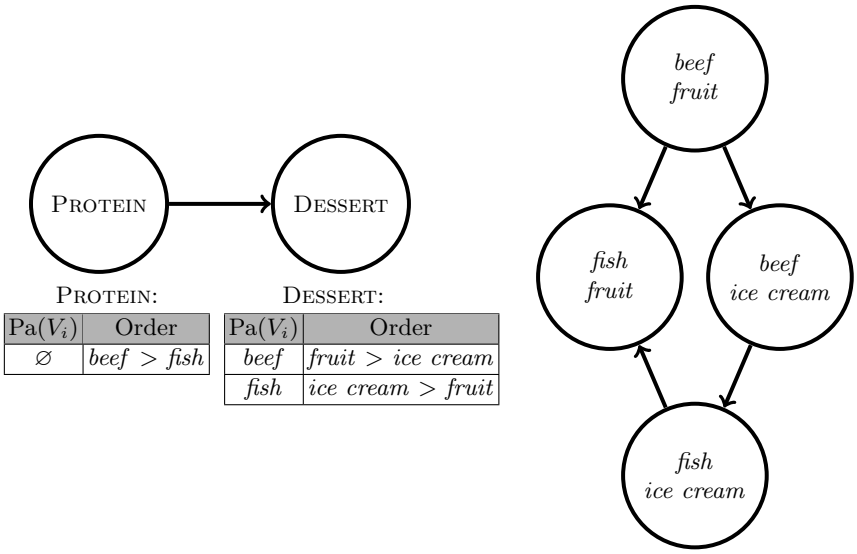
their decision making introspectively likely interferes with the very process we are studying, making introspection an ineffective method for eliciting preferences or thought processes [50,51,70]. Indeed, without actively allocating cognitive resources to commit information to memory, there is no reason to expect that a decision maker can accurately recall the deliberations underlying his decision afterwards. This is why psychologists run laboratory experiments where human actions in a controlled environment are observed, rather than asking people how they think. They draw inferences about latent preferences from observable quantities such as choice proportions, buying or selling prices, reaction times, eye movements, all of which need not reveal one single consistent picture [39,71]. The challenge is to model the relationship between theoretical constructs (e.g., preferences) and observed data (e.g., choices) [59].

2 Preferences in Computer Science

Preference handling in artificial intelligence is a robust and well developed discipline with its own working groups and specialized workshops [23]. Often, much of the work takes the form of creating or defining models and then analyzing the computational complexity of various reasoning tasks within these models [8,19]. One such model that has gained prominence since its introduction in 2004 is the CP-net [6]. A CP-net is a formal model able to capture conditional preference statements (cp-statements) such as, "For dinner, if I have beef, I prefer fruit to ice cream for dessert, but if I have fish, I prefer ice cream to fruit for dessert" and "I prefer beef over fish for dinner."

Formally, a CP-net [6] consists of a directed graph $G = \langle V, E \rangle$, where the nodes V represent *variables* (sometimes called *features*) of an object, each with its own finite domain or set of *values*. For each variable V_i in the graph there is a possibly empty set of parent variables $Pa(V_i)$. For each variable, an ordinal preference relation over its values is specified by a collection of cp-statements, called a *conditional preference table (CPT)*. The assignment of values to $Pa(V_i)$ can affect the preference relation over V_i. For example, in Figure 1 the variable PROTEIN can take the values *beef* or *fish* and *beef* is preferred. As PROTEIN has no parents, there is only one cp-statement. However, $Pa(\text{DESSERT}) = \text{PROTEIN}$ and therefore, depending on the assignment to PROTEIN either *fruit* is preferred to *ice cream* or vice versa.

A CP-net is a compact representation of the preference graph on *outcomes*. An outcome is a complete assignment of values to variables. The outcome graph $G_O = \langle V_O, E_O \rangle$ has nodes representing each possible set of values for the feature variables, and a directed edge between any two nodes that differ on exactly one feature value. The direction of the edge is determined by the preference over that feature, conditioned on the (otherwise fixed) values of its parent variables. The transitive closure of the preference graph gives the partial order over outcomes specified by the CP-net. A *sequence of worsening flips* is a directed path from an outcome o to o' through the outcome graph. This flipping sequence, if it exists, proves that outcome o is preferred to o'. We call this relation *dominance* and it



Protein:

$\mathrm{Pa}(V_i)$	Order
$\varnothing$	*beef* > *fish*

Dessert:

$\mathrm{Pa}(V_i)$	Order
beef	*fruit* > *ice cream*
fish	*ice cream* > *fruit*

Fig. 1. A simple CP-net (left) and the induced preference graph over outcomes (right).

is NP-hard to compute in the general case. Part of the difficulty of computing dominance arises because the outcome graph is exponentially larger than the CP-net graph, and improving flipping sequences can be be exponentially long [6]. Tractable subproblems exist (such as when the graph has the shape of a directed tree), as well as computational heuristics for determining dominance [34]. Understanding the CP-net as a human decision making tool may help us formalize other cases where reasoning with CP-nets is tractable, such as when the preference graph has low degree or the dominance relation has short flipping sequences [2].

Preference representation and reasoning plays a key role in many other areas broadly included under the umbrella of Artificial Intelligence (AI). For example, within the area of *constraint reasoning* [62], the annual Max-sat solver competition[6] includes problem instances that encode both hard constraints and soft preferences for domains such as scheduling, time-tabling, and facility location. Both traditional *social choice* and *computational social choice* [12] are fields that actively work with choice data and are beginning the transition towards working with repurposed data explicitly [44, 63]. However, these fields are still primarily focused on worst case assumptions, not behavioral or explanatory models of human decision making. Various forms of weighted logic representations such as *penalty logics* [17], *possibilistic logics* [20], and *answer set programs with soft constraints* [72] all explicitly rank states (assignments to all parameters) of the world. This ordered set of states is often interpreted as preferences over the states themselves. These fields primarily focus on algorithms for and quantifying the

[6] `http://maxsat.ia.udl.cat/introduction/`.

complexity of reasoning with choice data and preference models; testing these models and theories against human choice behavior is not a central focus.

The data focused fields of *machine learning* and *data mining* investigate preferences both implicitly and explicitly when working with, for example, large volumes of customer data [27]. This is perhaps most obvious in the sub-fields of *recommender systems* [61] and *preference learning* [22]. The objective in both of these areas is to learn and interpret observed choices (data) in order to make tangible recommendations or predictions, e.g. algorithms for the Netflix Prize Challenge [5] or Amazon product recommendations [35]. These areas are well developed preference handling fields where data are readily available from both academic sources [3] and as part of a number of industrial or commercial licenses or competitions (e.g. Kaggle, and the Yelp Academic Datasets). However, often these systems are only evaluated on their ability to minimize an error or loss function [5, 61] when compared to held out choice data (i.e., data not in the training set). The explicit goal of these systems is not to understand the features of a user's internal preference reasoning or if the system itself can affect the user's explicit choice.

Humans are being brought into the loop in more and more areas of computer science, often leading to important and exciting impacts. More researchers are focusing on understanding *how* users reason internally and *why* users implement recommendations [18, 56, 66]. Trust building through explanation in recommendation systems is becoming standard practice due to its increased effectiveness in leading to *implemented* recommendations [55]. Additional experiments with human subjects have also helped to validate activities like learning preferences through click tracking [28] and understanding the cognitive burden of asking certain preference queries [9]. The field of *computer-human interaction (CHI)* often performs studies of human behavior, validating models with laboratory experiment. Indeed, the most recent ACM Computer-Human Interaction conference (ACM:CHI 2014) provided courses on survey design and performing human studies [40, 47]. However, in the broad set of communities that deal with preferences in AI, the human element is still often misunderstood.

The omission of human centered testing bypasses both a host of practical considerations and formal verification of preference models. These problems require controlled human subjects experiments and offer exciting opportunities for cross disciplinary research. There are over 800 references to the original CP-nets paper with not a single human subjects study to investigate whether a CP-net is a model of human choice, nor any testing of the model for consistency with respect to the way individuals reason about individual preference.

3 Legal Considerations in Human Subjects Research

Collecting data from human participants involves a panoply of challenges including important legal and ethical considerations that are sometimes poorly understood or considered by researchers. We have encountered colleagues in the

US[7] and abroad from non-social science disciplines who had unknowingly broken laws by illegally gathering data from humans without undergoing appropriate prior review by an *Institutional Review Board (IRB)* and without undergoing legally required *ethics training*. While the IRB process can be cumbersome, it is an important step in using human subjects data. Modern tools such as Amazon's Mechanical Turk are a key resource [14, 43] that many in preference handling are embracing for collecting human subjects data [41].

Data from humans may or may not be considered *human subjects data*. Studying completely anonymized data sets is usually not considered *human subjects research*; hence, the ease of using data from a repository such as PrefLib [44] or the UCI Machine Learning Repository [3]. But if one can link data to the individual from whom those data came, then one operates under the strictures of human subjects research regulations. In many cases, however, an expedited process is in place when an IRB officer deems a study exempt, due to minimal risk, and waives the requirement of a full review by the board. Reviews by the board will evaluate a vast range of considerations or requirements, of which we review a few.

Incentivization: Generally, research in experimental psychology rewards participants in one of two major ways. Most experiments recruit undergraduate psychology students in exchange for course credit, this is commonly referred to as the "subject pool." Other experiments pay participants with cash or other rewards. Decision making experiments in this category often give some of the chosen options as real rewards in order to motivate participants to invest cognitive effort and reveal true preferences. Generally, behavioural and experimental economists disregard studies that do not link rewards to performance as being "insufficiently incentivized" [29]. There are also considerations of "over-incentivization" in that very large rewards can be blocked by some IRBs for being coercive. Another consideration is whether participants are allowed to receive payments, e.g., based on age, legal, or immigration status.

Informed Consent & Deception: Since the infamous "Milgram experiments" [45] in which participants were led to believe that they were torturing others, ethical issues in human subjects research have been discussed in great detail. Many protections have been put in place to protect participants in psychology, economics, and medical experiments from being harmed. Scientists and lab personnel are required to undergo extensive training, e.g., the *Collaborative Institutional Training Initiative (CITI,* https://www.citiprogram.org/), before they may carry out research on humans. Ethical issues of informed consent emerged prominently in the mass media recently when it came to light that Facebook carried out social and emotional experiments on some of its users without clear-cut informed consent [68]. In behavioural and experimental economics, outright deception is often frowned upon [29].

[7] Human-subjects standards vary from country to country and are also sometimes imposed by international academic societies.

CONFIDENTIALITY: It is straightforward that the protection of human subjects, besides avoiding immediate bodily or psychological harm, starts with proper confidentiality assurances. Considerations of what constitutes "anonymized" data is a growing concern in computer science and other disciplines. High-profile cases in recent years have shown that even a sequence of movie rental dates can be enough to discern personally identifiable information from a supposedly anonymized dataset [49]. Besides the obvious concerns about data trails from scheduling participants, time-stamped electronic data collection, and accounting records of payments, the use of cloud-based tools, such as storage or email, where servers may reside outside the country, or with commercial providers, threatens confidentiality.

4 Perspectives from Mathematical Psychology

Let $\mathcal{C}$ be a finite set of choice alternatives, and let $\succ$ denote pairwise preference, i.e., $x \succ y$ with $x, y \in \mathcal{C}$ denotes that a person strictly prefers x to y. Many models of preferences, including CP-nets, require $\succ$ to be transitive.

How would one test whether decision makers' preferences are transitive? Psychologists differentiate between *theoretical constructs* and *observables*. A binary preference relation, a real valued utility function, a CP-net, are theoretical constructs that we cannot observe directly, just as a physicist cannot observe gravity itself. In decision making, actual choices made by actual people are observables that are presumably related to the latent construct, just as an apple falling is an observable manifestation of gravity.

A major conceptual, mathematical, and experimental challenge for testing theories about preferences comes from the fact that decision makers experience uncertainty in what to choose when faced with multi-attribute options in which attributes trade off in complex ways. Experimentally, we observe substantial amounts of variability between people and even within a single person over repeated choices among the same options. It is not uncommon for a decision maker to choose x over y on 70% of occasions, and y over x otherwise, even within a one-hour study. This led economists and psychologists to mathematically model uncertainty and variability in choice. Arguably, the most natural way to model uncertainty in choice is via probabilistic models [4, 10, 36–38, 67].

There are two major classes of probabilistic choice models. One assumes that the theoretical construct of preference is deterministic but choices are probabilistic, the other assumes that the theoretical construct itself is probabilistic. For transitivity, the first model type assumes that each decision maker has one fixed deterministic preference $\succ$ over the course of the experiment, whereas the latter model casts preferences as a probability distribution over a set of transitive preferences. For CP-nets the analogue is to distinguish two major possibilities:

1. the decision maker uses one single fixed CP-net, but makes probabilistic *errors* in revealing this CP-net in overt choices;
2. the choice probabilities are *induced* by an unknown probability distribution over a collection of CP-nets.

An *error* specification may assume that the decision maker has unknown preference $\succ$, and if $x \succ y$ then she is more likely to pick x than y, formally and more precisely,

$$P_{xy} > \tau, \text{ with a bound on error rates of } 1 - \tau \leq \frac{1}{2}. \tag{1}$$

A *random preference specification* considers a (finite) collection $\mathcal{R}$ of permissible preference relations (e.g., transitive relations, CP-nets, etc.), a probability distribution $\mathbb{P}$ on $\mathcal{R}$, and models the binary choice probability P_{xy} as a marginal probability

$$P_{xy} = \sum_{\succ \in \mathcal{R}} \mathbb{P}(x \succ y). \tag{2}$$

Characterizing the binary choice probabilities that are consistent with a *random preference specification (2)* can be mathematically and computationally prohibitive. In the case that $\mathcal{R}$ is the collection of all strict linear orders over a finite set $\mathcal{C}$, the binary choice probabilities (2) form a convex polytope known as the *linear ordering polytope* [21,24,33]. The mathematical structure of this polytope is known only for small sizes of $\mathcal{C}$ and finding a complete minimal description in terms of *facet-defining inequalities* is computationally hard [42]. Building a random preference model in which the collection of permissible preferences $\mathcal{R}$ is composed of CP-nets would require that we understand the permissible binary choice probabilities (2). The currently standard approach would be to employ methods from polyhedral combinatorics, by defining and studying appropriate *CP-net-polytopes*, in which probability distributions over CP-nets are conceptualized as convex combinations of deterministic CP-nets.

We have sketched the conceptual and mathematical challenge of defining uncertain choices induced by theoretical preferences that form CP-nets, using probabilities. The next challenge is that those probabilities $\{P_{xy} \,|\, x \neq y; x, y \in \mathcal{C}\}$, in turn, are theoretical constructs. If we are to study CP-nets in the laboratory and if we are to allow different decision makers to use CP-nets differently, then we need to draw inferences about probabilities from finite samples using appropriate statistical tools. Both the error models (1) and the random preference models (2) impose multiple simultaneous order-constraints on the parameters of joint Bernoulli processes. This causes serious challenges in maximum-likelihood methods because point estimates may lie on the boundary of the parameter space (e.g., on a face of a convex polytope) where standard likelihood theory breaks down. Frequentist and Bayesian order-constrained likelihood-based inference methods have only become available recently [15, 32, 48, 59]. Some of the algorithms, e.g., for computing Bayes Factors between two competing convex polytopes, are computationally expensive, with current researchers sometimes using thousands of CPU-hours per Bayes Factor.[8]

[8] For an example of the complexities involved in testing transitivity of preferences, including a critical review of the prior literature, see, e.g. [11, 57, 58, 60].

The task, then, for a quantitative test of CP-nets in individual decision makers, includes: the development of "probabilistic specifications" that represent the uncertainty experienced by the decision maker, the adoption of suitable statistical tools, and the design and implementation of an experiment that generates data suitable for either testing the mathematical model as a hypothesis or for selecting between the model for CP-nets and alternative theoretical proposals. Both the mathematical characterization and the statistical inference involve significant mathematical and computational challenges.

5 Other Considerations for Laboratory Experiments

In defining a laboratory experiment on human decision making, attention must also be give to the following issues which, while not legal or ethical in nature, can affect the design and implementation of an experiment.

Data bias: Statistical inferences from finite sample data generally require repeated observations either from multiple people or from a given participant. In order to eliminate potential biases and the effect of irrelevant variables, a decision-making experiment asking participants to decide among choice options can implement a variety of "cross-balancing" precautions. These include, e.g., showing a given choice option randomly in different locations on a display to compensate for attentional biases and making different stimuli "equally complex" to balance cognitive load. Statistical tests and analyses often assume independent and identically distributed observations. These assumptions affect the experimental design itself, e.g., separating repeated observations through decoys to attenuate violations of independence.

Correlation vs. causality: This has important implications for selecting experimental methods over data mining or other approaches. If one wishes to make causal attributions that values in one variable "cause" outcomes in another variable, one needs to use random assignment to experimental conditions (e.g., placebo versus treatment).

Falsifiability, diagnosticity, and parsimony: According to these principles, theoretical predictions motivate what stimuli to use and hence precede data collection. Epistemologically, restrictive theories are favored because they lead to falsifiable predictions [54]. There are at least three major ways in which behavioral scientists use statistical inference.

1. Many scholars support a theoretical claim by statistically rejecting a *null hypothesis* of "no effect," a practice that has come under intense criticism [13, 46, 69].
2. Others, similar to data mining methods, formulate mathematical models and use statistics to estimate parameters through data fitting, then interpret the inferred parameter values in terms of scientific primitives. Oftentimes the validity or replicability of the findings are assessed through goodness-of-fit on hold-out samples or through predictions about future data.

3. More and more behavioral scientists use Bayesian methods to carry out competitions among theories that vary in their parsimony, by weighing prior beliefs with empirical evidence, and penalizing flexible models [31, 48].

Several disciplines within social science are currently engaged in a major debate about replicability,[9] publication bias, and scientific integrity [26, 64, 65]. Most social science journals only consider novel findings for publication, leading some researchers to draw scientific conclusions from very slight statistical effects, and several high-profile scholars have been accused and/or found guilty of faking their data. The practical consequence of the recent debate is that researchers must take care to ensure that their models and experiments stem from rigorous theories, which make precise predictions that can be tested in a laboratory setting through the use of appropriately applied statistics.

6 Case Study: The CP-net Experiment

We have recently completed data collection on an experiment to test whether decision makers subjectively represent preferences in a way that is consistent with a mathematical CP-net representation. We have incorporated the considerations above with many additional practical and logistic constraints.

A good rule of thumb for running a first experiment in a given domain is to start simple. Since there is no prior empirical work on actual CP-nets of actual people, we needed to design the study without having to hypothesize too many details about CP-nets that are suitable for the domain under consideration. Otherwise, were we to conclude that "our" CP-net is not descriptive of our participants, we would not learn much about the general descriptive validity of CP-nets. If we allow *all* CP-nets on a given set of choice options as potential preference states, then we need to limit the number of CP-nets that are possible. We do not want to make the CP-nets trivial but we also cannot make them overly complex as it will lead to intractable experiments. Therefore we limit ourselves to acyclic dependency graphs with four binary nodes/variables. This means that our CP-nets have 16 choice alternatives. This permits a rich set of preference states, exactly 481,776 distinct CP-nets (computed as all possible non-degenerate boolean functions on $n = 4$ binary variables [1]).

The next major set of considerations is to decide on actual stimuli that are both interesting and that may tell us something about everyday decision making, at least at face value. Furthermore, at least some of the stimuli need to be 'deliverable' as real prizes while the other ones need to be 'cross-balanced.' We therefore selected two domains, restaurant menu choices (since they are common hypothetical illustrations in the CP-net literature) and choices among retail goods or services. For example, for the restaurant menu options we chose "appetizer" versus "dessert" as one attribute, "chicken" versus "shrimp" as another attribute, etc.

[9] See, e.g., http://psychfiledrawer.org/.

Since we incentivized our participants by offering them some of their choices as real rewards, team members spent significant time contacting retail and restaurant managers to find ways to purchase rewards through university purchase orders and to ensure that a person will be given the exact reward we specify (as opposed to being able to use, say, a gift certificate in a fungible way). Likewise, multiple team members agonized over finding a sufficiently rich set of 'comparable' stimuli, even for trials that are not used to determine real rewards. For example, all stimuli need to be credible as potential rewards of a comparable value and payable by a federal grant. Over two domains with 60 participants the experiment distributed rewards of $4992.00 USD. The distributed rewards consisted of $2400.00 of prescribed meals at a local restaurant, $220.00 of video rentals, and $2372.00 of merchandise at the university bookstore.

There are also major tradeoffs between practical, logistical and statistical prerogatives: Ultimately, participants need to make sufficiently many choices among sufficiently many options to allow statistical estimation, hypothesis testing, or model selection. We decided to make each "trial" of the experiment a *ternary paired comparison*, i.e., two meals are presented and the decision maker can either express a preference for one, the other, or express "no preference." Statistically, this means that each "trial" provides an observation for a trinomial random variable. In order to obtain repeated observations, we needed to show each pair several times, at the risk of making the experiment laborious and repetitive. Hence, we substituted different "instantiations" of a given "choice" on different trials, "chicken" on one trial could be "Chicken Marsala" and on another trial could be "Chicken della Nonna." However, this means that we may have introducing many unintended variables that we are not modeled in the CP-net. Therefore, when showing a participant two meals that share the value "chicken" for the variable "main dish," we showed them the identical chicken dish in both options, so as to make it impossible for them to have a pairwise preference on a variable that we model as having identical values in both options.

There is a tradeoff between the number of times we ask a user to make a decision and the statistical tests we can then employ to perform reliable statistical analysis. Since we are asking users to choose between two meals, with 16 total meals, that gives 120 trials or pairwise comparisons that we must elicit from each user, and each of these trials must be repeated. Some of our models are convex (polytopes), in which case we can pool data across subjects even if there are individual differences between them. Some of our (error) models are not convex and should best be evaluated separately for each individual. Advanced statistical methods that do not require asymptotic statistics can get by with fewer than 10 trials per user per question.

The tasks involved in preparing for this experiment include: computing a list of all 4-variable CP-nets; developing the initial set of variables for each domain; negotiating agreements with the Institutional Review Board for human-subject research; getting agreement from vendors to provide specified rewards (and dealing with the video rental business going out of business before some of the long term rewards could be redeemed); creating multiple equivalent wordings

of the same reward (e.g., 6 T-shirts in one trial and a half dozen short sleeved shirts in another); developing and testing the GUIs and interface functionality on iPads for the experiment. These tasks took about 350–400 person hours. As members of the team are extremely experienced with navigating the bureaucracy of IRB approval and negotiating non-fungible rewards with outside vendors, this number likely underestimates the time required for a first time experimenter.

For data collection, we consider both the participants' time and the cost of running the study: 2 sessions for each of the 60 participants and about 90 minutes per session. There are 5 iPads, but scheduling is complicated, so we ran about 50 experimental sessions to get the data from those 60 participants. The person overseeing each experimental session spent about an hour for each experimental session distributing and collecting the informed consent paperwork, making sure the app was running on the iPads and ready to use, making sure the iPads were charged, introducing people to the study, answering questions, explaining the payment, scheduling their second session, making sure the results were uploaded, making sure each participant's payment was provided confidentially in a separate room, making sure the post-test questionnaire was filled out, etc. Thus, the number of person-hours for running the experiment (about 180 person-hours of subjects' time, plus about 100 hours of experimenters' time) was slightly less than the time spent preparing the experiment itself.

7 Conclusion

In this paper we have highlighted some of the key pitfalls and challenges associated with human subjects experiments within preference model testing. Ideally, experimentalists in AI can use this as both a call to action and as a starting point for conducting their own experiments both in human subjects labs and leveraging the power of online tools such as Mechanical Turk [41, 43]. We have only skimmed the surface of the relevant literatures in computer science and psychology. There is a vast literature on experimental studies in other fields including decision sciences, experimental economics, medical, and other cognitive studies areas. We hope this article serves as a jumping off point into the literature.

Data are available in large quantities, but we should resist the temptation to rely on past data alone when testing a preference modeling framework. Human experimentation should be part of the testing process. However, in doing this, we need to pay attention to several conceptual, mathematical, statistical, computational, legal and ethical considerations, as well as tackle many practical and logistic complications. We believe that AI and psychology researchers should work together in this endeavor. For AI researchers, understanding the functions and limitations of human decision making can lead to the development of more accurate models and heuristics in the multitude of areas that engage with humans and preferences. For psychologists, understanding the computational burden of reasoning with various preference models can inform new experiments and processes.

We have just completed data collection at the time of acceptance of this manuscript, after clearing all the significant development and logistical hurdles we have outlined in this paper. Proper analysis of this data will take months; discrepancies in the publication culture of computer science and psychology means we must target psychology journal submissions first (as data must be novel for publication). This will give us the first real experiment which will contemplate the question of whether or not subjects' preferences over two domains (retail and food) are at least noisily consistent with CP-net models and whether or not, given adequate instruction, the subjects can write these preferences down in a way that is consistent with their previous choices.

References

1. Allen, T.E., Goldsmith, J., Mattei, N.: Counting, ranking, and randomly generating CP-nets. In: 8th Workshop on Advances in Preference Handling (MPREF 2014), AAAI-14 Workshop Series (2014)
2. Allen, T.E.: CP-nets with indifference. In: Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on. pp. 1488–1495. IEEE (2013)
3. Bache, K., Lichman, M.: UCI Machine Learning Repository (2013), `http://archive.ics.uci.edu/ml`, University of California, Irvine, School of Information and Computer Sciences.
4. Becker, G., DeGroot, M., Marschak, J.: Stochastic models of choice behavior. Behavioral Science 8, 41–55 (1963)
5. Bennett, J., Lanning, S.: The Netflix prize. In: Proc. KDD Cup and Workshop (2007)
6. Boutilier, C., Brafman, R., Domshlak, C., Hoos, H., Poole, D.: CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. Journal of Artificial Intelligence Research 21, 135–191 (2004)
7. Boutilier, C., Patrascu, R., Poupart, P., Schuurmans, D.: Constraint-based optimization and utility elicitation using the minimax decision criterion. Artificial Intelligence 170(8), 686–713 (2006)
8. Brafman, R.I., Domshlak, C.: Preference handling—an introductory tutorial. AI Magazine 30(1), 58 (2009)
9. Braziunas, D., Boutilier, C.: Assessing regret-based preference elicitation with the utpref recommendation system. In: Proceedings of the 11th ACM Conference on Electronic Commerce (EC). pp. 219–228. ACM (2010)
10. Carbone, E., Hey, J.: Which error story is best? Journal of Risk and Uncertainty 20, 161–176 (2000)
11. Cavagnaro, D., Davis-Stober, C.: Transitive in our preferences, but transitive in different ways: An analysis of choice variability. Decision 1(1), 102–122 (2014)
12. Chevaleyre, Y., Endriss, U., Lang, J., Maudet, N.: Preference handling in combinatorial domains: From AI to social choice. AI Magazine 29(4), 37–46 (2008)
13. Cohen, J.: The earth is round ($p < .05$). American Psychologist 49(12), 997 (1994)
14. Crump, M.J.C., McDonnell, J.V., Gureckis, T.M.: Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PloS one 8(3), e57410 (2013)

15. Davis-Stober, C.: Analysis of multinomial models under inequality constraints: Applications to measurement theory. Journal of Mathematical Psychology 53, 1–13 (2009)
16. De Groot, A.D.: Thought and Choice in Chess (Psychological Studies), vol. 4. Mouton de Gruyter, 2nd edn. (1978)
17. De Saint-Cyr, F.D., Lang, J., Schiex, T.: Penalty logic and its link with Dempster-Shafer theory. In: Proc. UAI. pp. 204–211 (1994)
18. Dodson, T., Mattei, N., Guerin, J., Goldsmith, J.: An English-language argumentation interface for explanation generation with Markov decision processes in the domain of academic advising. ACM Transactions on Interactive Intelligent Systems (TiiS) 3(3), 18 (2013)
19. Domshlak, C., Hüllermeier, E., Kaci, S., Prade, H.: Preferences in AI: An overview. Artificial Intelligence 175(7), 1037–1052 (2011)
20. Dubois, D., Lang, J., Prade, H.: A brief overview of possibilistic logic. In: Symbolic and Quantitative Approaches to Uncertainty, pp. 53–57. Springer (1991)
21. Fiorini, S.: Determining the automorphism group of the linear ordering polytope. Discrete Applied Mathematics 112, 121–128 (2001)
22. Fürnkranz, J., Hüllermeier, E.: Preference Learning. Springer (2010)
23. Goldsmith, J., Junker, U.: Preference handling for artificial intelligence. AI Magazine 29(4), 9 (2009)
24. Grötschel, M., Jünger, M., Reinelt, G.: Facets of the linear ordering polytope. Mathematical Programming 33, 43–60 (1985)
25. Guo, S., Sanner, S., Bonilla, E.V.: Gaussian process preference elicitation. In: Advances in Neural Information Processing Systems. pp. 262–270 (2010)
26. Ioannidis, J.: Why most published research findings are false. PLoS Medicine 2(8), e124 (2005)
27. Jawaheer, G., Weller, P., Kostkova, P.: Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. ACM Transactions on Interactive Intelligent Systems (TiiS) 4(2) (June 2014), `http://doi.acm.org/10.1145/2512208`
28. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., Gay, G.: Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Transactions on Information Systems (TOIS) 25(2), 7 (2007)
29. Kagel, J., Roth, A.: The Handbook of Experimental Economics. Princeton University (1995)
30. Kamishima, T.: Nantonac collaborative filtering: Recommendation based on order responses. In: The 9th International Conference on Knowledge Discovery and Data Mining (KDD). pp. 583–588 (2003)
31. Kass, R., Raftery, A.: Bayes factors. Journal of the American Statistical Association 90(430), 773–795 (1995)
32. Klugkist, I., Hoijtink, H.: The Bayes factor for inequality and about equality constrained models. Computational Statistics & Data Analysis 51, 6367–6379 (2007)
33. Koppen, M.: Random utility representation of binary choice probabilities: Critical graphs yielding critical necessary conditions. Journal of Mathematical Psychology 39, 21–39 (1995)
34. Li, M., Vo, Q.B., Kowalczyk, R.: Efficient heuristic approach to dominance testing in CP-nets. In: Proc. AAMAS. pp. 353–360. Richland, SC, USA (2011)
35. Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. Internet Computing, IEEE 7(1), 76–80 (2003)
36. Loomes, G., Sugden, R.: Testing different stochastic specifications of risky choice. Economica 65, 581–598 (1998)

37. Luce, R.: Individual Choice Behavior: A Theoretical Analysis. John Wiley, New York (1959)
38. Luce, R.: Four tensions concerning mathematical modeling in psychology. Annual Review of Psychology 46, 1 – 26 (1995)
39. Luce, R.: Joint receipt and certainty equivalents of gambles. Journal of Mathematical Psychology 39, 73–81 (1995)
40. MacKenzie, I.S., Castellucci, S.J.: Empirical research methods for human-computer interaction. In: Proc. CHI. pp. 1013–1014 (2014)
41. Mao, A., Procaccia, A.D., Chen, Y.: Better human computation through principled voting. In: Proc. 27th AAAI Conference on Artificial Intelligence (AAAI) (2013)
42. Martí, R., Reinelt, G.: The Linear Ordering Problem: Exact and Heuristic Methods in Combinatorial Optimization, vol. Applied Mathematical Science 175. Springer (2011)
43. Mason, W., Suri, S.: Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods 44(1), 1–23 (2012)
44. Mattei, N., Walsh, T.: Preflib: A library for preferences, http://www.preflib.org. In: Proc. ADT. pp. 259–270 (2013)
45. Milgram, S.: Behavioral study of obedience. The Journal of Abnormal and Social Psychology 67(4), 371 (1963)
46. Morey, R., Rouder, J., Verhagen, J., Wagenmakers, E.: Why hypothesis tests are essential for psychological science: A comment on Cumming. Psychological Science 25(6), 1289–1290 (2014)
47. Müller, H., Sedley, A., Ferrall-Nunge, E.: Designing unbiased surveys for HCI research. In: Proc. CHI. pp. 1027–1028 (2014)
48. Myung, J., Karabatsos, G., Iverson, G.: A Bayesian approach to testing decision making axioms. Journal of Mathematical Psychology 49, 205–225 (2005)
49. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: IEEE Symposium on Security and Privacy. pp. 111–125 (2008)
50. Nisbett, R.E., Wilson, T.D.: Telling more than we can know: Verbal reports on mental processes. Psychological review 84(3), 231 (1977)
51. Nordgren, L.F., Dijksterhuis, A.: The devil is in the deliberation: thinking too much reduces preference consistency. Journal of Consumer Research 36(1), 39–46 (2009)
52. Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W.P., Jonker, C.M.: Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. User Modeling and User-Adapted Interaction 22(4-5), 357–397 (2012)
53. Popova, A., Regenwetter, M., Mattei, N.: A behavioral perspective on social choice. Annals of Mathematics and Artificial Intelligence 68(1-3), 5–30 (2013)
54. Popper, K.: The Logic of Scientific Discovery. London: Hutchinson (1959)
55. Pu, P., Chen, L.: Trust building with explanation interfaces. In: Proc. of the 11th International Conference on Intelligent User Interfaces (IUI). pp. 93–100 (2006)
56. Pu, P., Chen, L., Hu, R.: Evaluating recommender systems from the user's perspective: survey of the state of the art. User Modeling and User-Adapted Interaction 22(4-5), 317–355 (2012)
57. Regenwetter, M., Dana, J., Davis-Stober, C.P.: Transitivity of preferences. Psychological Review 118, 42–56 (2011)
58. Regenwetter, M., Dana, J., Davis-Stober, C.: Testing transitivity of preferences on two-alternative forced choice data. Frontiers in Quantitative Psychology and Measurement (2010)

59. Regenwetter, M., Davis-Stober, C.P., Lim, S.H., Guo, Y., Popova, A., Zwilling, C., Cha, Y.C., Messner, W.: QTest: quantitative testing of theories of binary choice. Decision 1(1), 2–34 (2014)
60. Regenwetter, M., Davis-Stober, C.: Behavioral variability of choices versus structural inconsistency of preferences. Psychological Reviewe 119(2), 408–416 (2012)
61. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.): Recommender Systems Handbook. Springer (2011)
62. Rossi, F., Beek, P.V., Walsh, T.: Handbook of Constraint Programming. Elsevier (2006)
63. Rossi, F., Venable, K.B., Walsh, T.: A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice. Synthesis Lectures on Artificial Intelligence and Machine Learning 5(4), 1–102 (2011)
64. Schooler, J.: Unpublished results hide the decline effect. Nature 470(7335), 437 (2011)
65. Simmons, J., Nelson, L., Simonsohn, U.: False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22(11), 1359–1366 (2011)
66. Tintarev, N., Masthoff, J.: Designing and evaluating explanations for recommender systems. In: Recommender Systems Handbook, pp. 479–510. Springer (2011)
67. Tversky, A.: Intransitivity of preferences. Psychological Review 76, 31–48 (1969)
68. Waldman, K.: Facebook's unethical experiment. Slate (2014), http://www.slate.com/articles/health_and_science/science/2014/06/facebook_unethical_experiment_it_made_news_feeds_happier_or_sadder_to_manipulate.html
69. Wetzels, R., Matzke, D., Lee, M., Rouder, J., Iverson, G., Wagenmakers, E.: Statistical evidence in experimental psychology an empirical comparison using 855 t tests. Perspectives on Psychological Science 6(3), 291–298 (2011)
70. Wilson, T.D., Schooler, J.W.: Thinking too much: introspection can reduce the quality of preferences and decisions. Journal of personality and social psychology 60(2), 181 (1991)
71. von Winterfeldt, D., Chung, N.K., Luce, R., Cho, Y.: Tests of consequence monotonicity in decision making under uncertainty. Journal of Experimental Psychology: Learning, Memory and Cognition 23, 406–426 (1997)
72. Zhu, Y., Truszczynski, M.: On optimal solutions of answer set optimization problems. In: Logic Programming and Nonmonotonic Reasoning, pp. 556–568. Springer (2013)