Climate control for performant HPUs

April 10, 2014

Abstract

Introduction

There are still many tasks for which humans outperform computers These tend to be tasks that rely on a large repertoire of prior knowledge, the excercise of common sense judgment, certain visual tasks, and tasks that require spontaneous hypothesis generation.

With the emergence of microtask platforms like Amazon Mechanical Turk (AMT), comes the hope of building compute in which human intelligence can be accessed on-demand. This leads to an analogy of the human to the CPU, wherein we think of human processing units.

A major challenge in building systems that deliver on this vision is the natual variability obtained in HPU output. In some tasks, this variability is desireable. For example, in an image labelling task, the natural variability in HPU output leads to greater coverage of the semantic space occupied by an image. In other cases this variability can be problematic, such as in a transcription task, where there is only one corect output, and any variability is only noise.

Whether desired or not, to build up efficient systems from HPUs, one must understand the natural variance quantitatively, and understand the factors that influence it.

Here we model the variance between HPUs as comming from two sources. The first is a persistent, or intrinsic variability. We imagine that this to encompass temperamental predispositions, life history, and developmental stage of the individual. While it may change over the course of a person's lifetime, in the view of the computational architect, it is essentially constant.

The seccond is a short-term variability, which arises from the fact that peoples immediate state is a function of recent events. This would include such things as an HPU becomming more efficient at a task once it has completed a few tasks of the same type. It would also include such things as the potentially detrimental effects of a noisy environment.

In this view, one can imagine HPUs being different in their specifications, and also exhibiting hysteresis, whereby their output at some time t is a function of their inputs at time t, $t-1$, and so on.

In psychology, the act of producing these short-lived states that effect a person's performance during a task is called *priming*. We will adopt this term throughout the present work.

Prior Work

Priming itself can be broken into many types, and many studies exist exploring the effects of overt priming on the performance of HPUs. Priming may arise due to how the task is framed. This includes such things as the stated purpose of the task, and the identity of the requester the task.

In [1], the researchers investigate the effects of framing task either in a meaningful or meaningless way. Compared to a zero-context control treatment, workers increase their output (for less pay), when they are told that they are helping identify cancer, but there is no change in quality. When workers are told that their submissions will be discarded, there is no change in the amount of work done, but the quality declines.

Another source of priming can be due to what might be called *sidestream information*: information that is presented to the individual but which is not actually salient to the task.

Other researchers have studied how having peer's responses available to the task could influence the worker.

In the present work, we focus on a more subtle source of priming, which arises simply from the worker's performance in earlier stages of the task.

We compare this to priming arising from framing the task by disclosing a (fictitious) funding agency that is paying for the study in which they are participating.

Framework

At this point we have surveyed some of the conceptual framework traditionally used to discuss priming in the context of psychology. Having the origin that they do, these frameworks are suited to describing the internal mechanisms of the mind from which they arise. For our purposes here, we are interested primarily in the *effects* of priming, and so it makes sense to adopt a position conducive to those ends. Since our interest in priming arises from a desire to construct computing systems from HPUs, we shall propose a well-defined algorithmic definition for priming.

We must first remark that it does not make sense to speak of an un-primed state. A person can be no more un-primed than a ship at sea can have no bearing. Next, we propose that priming should generally be regarded as a property of a population of HPUs, rather than as the state of a single HPU. This is rather like the notion of temperature in thermodynamics. Although it does make sense to speak of the kinetic energy of a single molecule, Temperature is a statement about the distribution of kinetic energies in a population of molecules. The definition of priming which we presently submit will also be defined in terms of populations.

Two populations, J and K are said to be *differently primed* with respect to a task $\mathcal{T}$ if there exists a polynomial-time algorithm $\mathcal{A}$ that can distinguish (classify) members of J and K with accuracy θ, when $\mathcal{A}$ is provided the labelled work-products of HPUs from J and K tasked with $\mathcal{T}$. When $\mathcal{A}$ exists, we say that J and K deviate by θ in priming.

The above definition is simply intended to provide a well-defined definition of what priming *is*. Naturally, it says nothing about the consequences of priming. The significance of the priming of a given population will depend both on the nature of $\mathcal{T}$ and on the intended purpose of the work products derived from HPUs performing $\mathcal{T}$.

Methods

Task set-up. We paid 900 AMT workers to perform an image-labelling task. A task consisted of labelling 10 images, with 5 labels each. The first 5 images were varied depending on the priming treatment, while the last 5 images were the same accross all treatments. Ordering of the images was kept constant.

Workers were randomly assigned to one of 6 treatments. The treatments differed from one another along two dimensions. The first dimension consisted of varying the first 5 images shown to the worker. This was used to test the effects of *in-task* priming.

The second dimension concerned disclosure of a (fictitious) organization, purportedly funding work as part of a research study. Depending on the treatment, one of two funding agencies was presented, or no indication was made.

Tasks were presented to workers as a series of panels or flash cards. The first panel provided brief instructions, and was identical for all treatments. Workers could see this panel when previewing the task, but could not advance. Depending on the treatment, the worker was either shown a second pannel stating the name of one of two fictitious organizations funding the work, or this panel was skipped. The next five pannels each consisting of a priming sub-tasks, wherein the worker was asked to submit five descriptive labels. The images used during the priming sub-task depended on the treatment. The last 5 pannels consisted of testing subtasks, wherein, as for the priming sub-tasks, workers were asked to submit 5 descriptive labels.

Choice of images. The 5 test images, were chosen with two ideals in mind. First, we chose images that we judged would generate a diverse vocabulary of labels, such that the effects of priming could be detected. In other words, sparse images with a single object in the foreground were not considered good candidates, since they would be less likely to elicit labels that varied from one worker and one priming treatment to the next.

Second, we chose images which would produce labels belonging to two broad concepts, which would serve as the targets of our priming: food and culture. This created the opportunity to attempt to prime workers in a way that would bias them toward emitting food-related or culture-related labels.

Under these considerations, we chose the images shown in Fig. 5. Each of these images has food as its main focus, but also has a strong and specific cultural reference due to the unique, iconic character of the food and the artifacts depicted.

To investigate in-task priming, we chose a set of images that highly recognizeable cultural settings and no food, and another set that contained separated food ingredients, without any overt cultural content. The third set of images was chosen to be very much like the test images, showing prepared meals, and though prepared food is inseperable from culture, these images were chosen based on being culturally more muted or ambiguous.

Label ontology. In order to provide a deeper analysis, we built an ontology of the corpus of all labels applied to the first test image. The ontology was built as a directed acyclic graph starting

Results and discussion

Priming affects HPU output. Before looking for differences in the content of labels provided by workers from different treatments, we first demonstrate that the treatments are distringuishable in an algorithmic sense.

Using a naive bayes classifier, we are able to distinguish with high precision and specificity between workers from the AMBG treatment and any of the other 5 treatments. Fig 1A shows F1-score for a naive bayes classifier when distinguishing between the AMBG and the other treatments. The classifier achieves high precision and specificity in task. Figure 1B shows the F1-score for binary classification between the CULT_{img} and the other treatments, while 1C shows that for INGR_{img} and the other treatments.

Not surprisingly, pairs of treatments in which one primes for culture, and the other primes for ingredients are easily distinguished. However, it is quite surprising that high accuracy is achieved when classifying treatments within the same orientation (e.g. cultural). This is especially true in such cases as the classification of CULT_{img} and $\text{CULT}_{fund,img}$ ($\text{F}_1 = 0.863$), where both treatments also share the priming images in common.

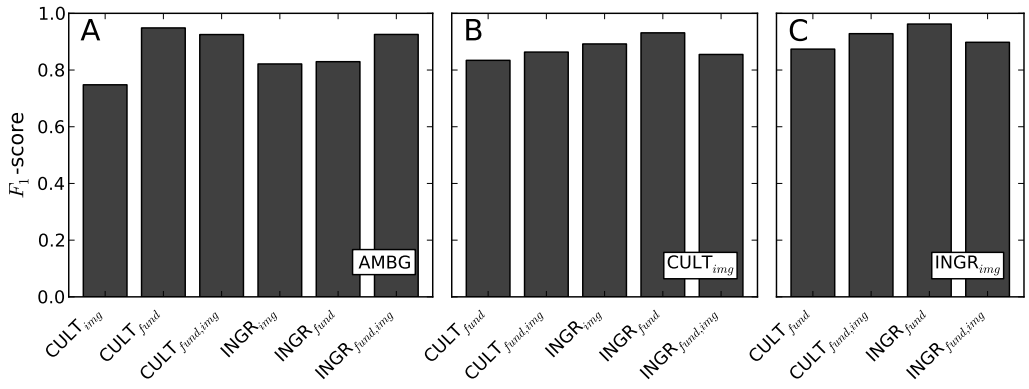

Figure 1: F_1-score for binary classification of HPUs from seperate treatments using a naive Bayes classifier. Each pannel shows the performance of the classifier when distinguishing between a basis treatment (inset) and the treatments listed on the abscissa.

We find it remarkable that, using only the labels that workers provide, it is possible to infer with good accuracy (at least when given a choice between two posibilities) the treatment to which the worker was subjected.

Priming orients HPU focus. Having established that each treatment does in fact consist of distinguishable populations of the HPUs, we next look at the content of labels. Two of the root labels in our ontology, `food` and `cultural`, map naturally onto the two concepts that laid behind the design of our priming treatments. A natural and straightforward expectation is that HPUs from treatments CULT$_x$ should emit more labels that are ontological descendents of `cultural`, while those from the INGR$_x$ treatments should emit more labels under `food`.

In Fig. 2, we exhibit the percentage of labels having a cultural orientation (i.e. descending from `cultural` in the ontology), and the percentage having a food orientation. Before making comparisons using this information, we remark that there is no reason to expect a balance in the number of words having cultural or food orientation overall. The AMBG treatment, which was designed to promote neither orientation, exhibits a significant fraction of words from both the food and cultural orientation, while significantly favoring the former.

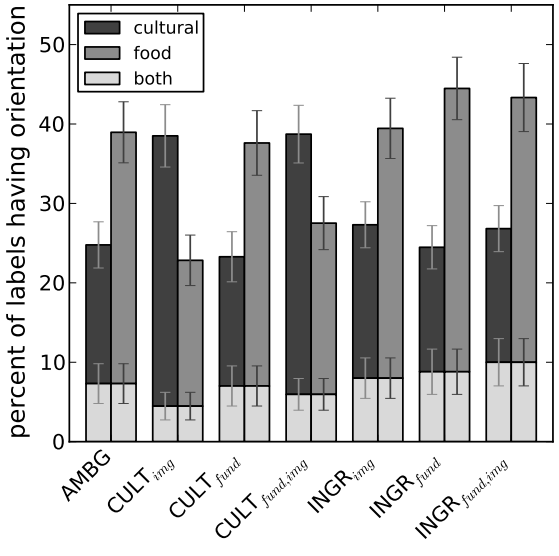

Figure 2: Percentage of labels of a food- or cultural-orientation, or both. In our ontology of labels, a label can have multiple parents. For example `naan` inherits from both `food` and `cultural` through its parent `indian food`.

Both CULT$_{img}$ and CULT$_{fund,img}$ show a significant excess of culturally-oriented labels and fewer food-orientel labels. Furthermore, HPUs in these treatments emit more culturally-oriented labels even while emitting fewer labes that are simultaneously oriented toward cultural and food (light bars in Fig. 2). The deviation of these treatments from the others is well beyond the 95% confidence interval.

Interestingly, the CULT$_{fund}$ treatment did not have this effect. Although we have demonstrated that CULT$_{fund}$ is differently primed from AMBG using a naive bayes classifier, in respect of overal fraction of food- and culturally-oriented labels, no distinction is to be made.

In the case of INGR$_x$ treatments one expects to see an enrichment of food-oriented labels. There is perhaps some evidence for this INGR$_{fund}$ and INGR$_{fund,img}$, but we cannat make any assertion with confidence.

Priming affects attention to detail. We next look at how alternately treated HPUs differ in their tendency to use more specific or more general labels. We use the ontology of labels emitted on the first test image to unambiguously establish a partial ordering of label-specificity. If one label ℓ_1 is within the ancestry of another ℓ_2, we say that ℓ_1 is more general than ℓ_2, otherwise they are not comparable. Thus, the label `naan` is more specific than both `bread` and `indian`, while uncomparable to `statue`.

Label-specificity only generates a partial ordering—at least in our experimental design. A consequence of this is that it is not possible to assign an overall specificity score to a set of labels. We submit that this reflects the underlying complexity of natural language semantics. While we admit that there may be many ways to generate full orderings based on some notion of label specificity, we believe that this would inappropriately collapse qualitative differences between labels, leading to results that are difficult to interpret.

In Fig. 3, We show pairwise specificity comparisons between various pairs of treatments. Figure 3A shows the comparison of AMBG with all other treatments. Both treatments primed with cultural images (CULT$_{img}$ and CULT$_{fund,img}$) exhibit an excess of more general words compared to AMBG. Meanwhile, INGR$_{fund}$ emitted an excess of more specific words.

We can seek to explain these observations by restricting the comparison to labels of a specific orientation (e.g. food or cultural). When we perform the same comparisons but restricting to food-oriented labels, as shown in Fig. 3G, we see some of the same tendencies as in A. However, when restricting to culturally-oriented labels (Fig 3D) there are no significant comparative deviations in specificity to speak of.

A first conclusion that we can draw from this is that there is a significant loss of specificity in labels emitted by treatments having non-ambiguous priming images. In Fig. 3G, only CULT$_{fund}$ and INGR$_{fund}$ show no significant deviation.

- through negative priming workers stop responding to generic features of images of prepared meals, which happens in the ambiguous case.

- On the other hand, workers presented with either the cultural or ingredients images are initially struck by the gross differences, which draws generic labels, especially with respect to food for which generic labels in ambiguous treatment had been suppressed through negative priming.

- it seems, at first, inconsistent that CULT$_{img}$ and CULT$_{fund,img}$ would at once enrich the fraction of culturally-oriented words, but produce no change in the specificity of culturally oriented words. First of all, there is no logical incompatibility with producing a greater number of cultural terms, yet which are generic in nature. If we follow the hypothesis that focus (or orientation) is directed by positive priming (perhaps through the subconscious or conscious inference of reqeuster intent), while focusing on nuances comes from the inhibition of gross fea-

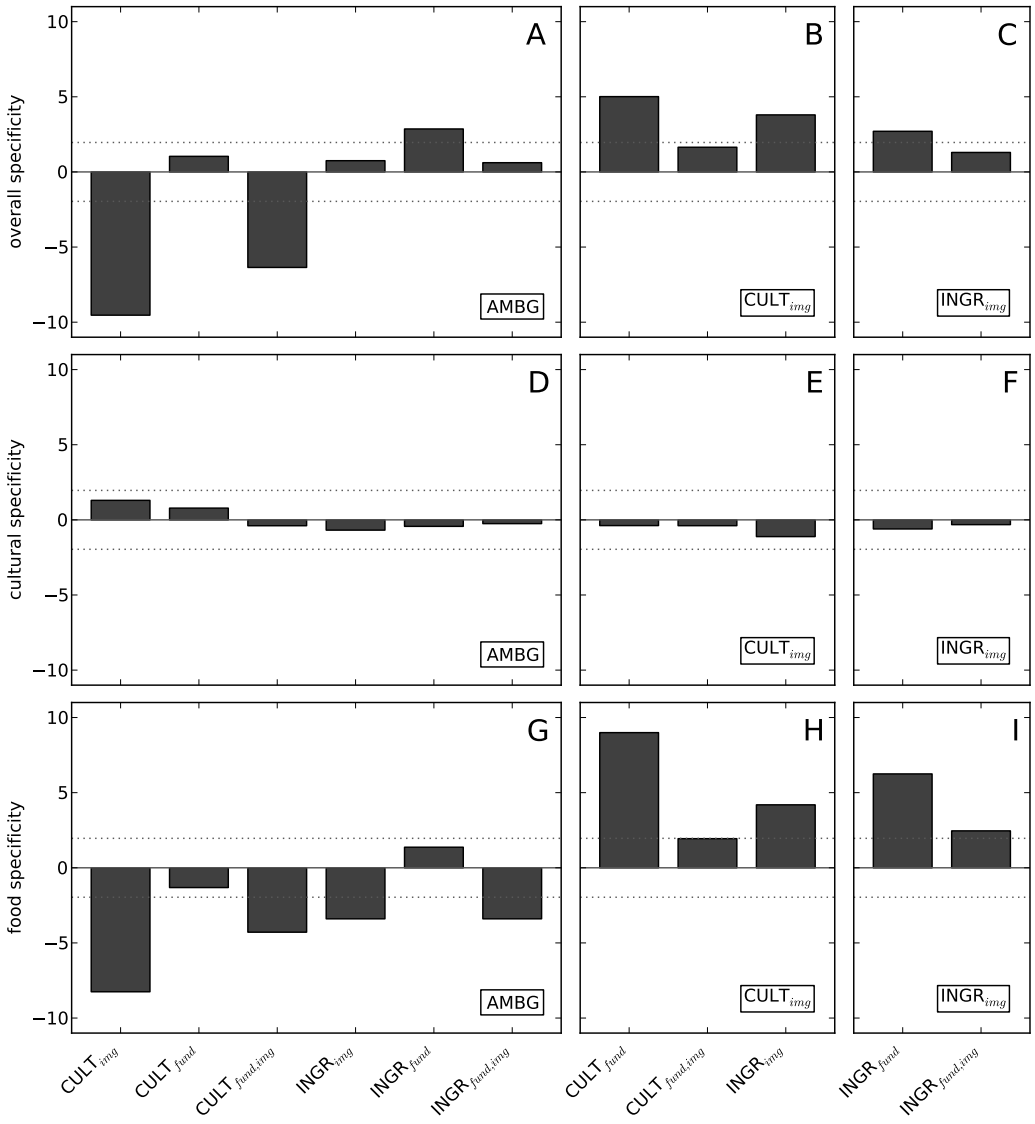

Figure 3: Pairwise comparisons of label specificity between different HPU treatments. Each pannel presents a binary comparison between a basis treatment (inset) and the subject treatments indicated on the abscissa. A positive specificity score indicates that the subject treatment emitted more specific words than the basis treatment overall. In a given comparison, a sample of 50 HPUs from the each treatment was randomly sampled, and the specificity of labels from the HPUs of oposing treatments were compared. To compare two HPUs, each pair of labels from different HPUs are compared. and the specificity score of subject HPU is the number of cases where its label is more specific than the subject HPUs, less the number of cases where it is more general. This score is avereaged for all HPU pairings. Statistical significance is guaged by generating a null-comparison between two mutually exclusive subsamples from the basis treatment. This null-comparison yields a distribution of relative specificities whose mean is in principle zero. The specificity scores are expressed in terms of standard deviations of the null-comparison specificities. The dotted lines represent the 95% confidence interval for rejecting the null Hypothesis that the basis treatment and subject treatment are equally specific.

tures through negative priming, then this combination of observations makes sense. Having seen many cultural images, HPUs in CULT_{img} and $\text{CULT}_{fund,img}$ are primed to emit cultural labels, however, since the diversity of images is high compared to other treatments on reaching the test images, there no opportunity for focussing in to finer detail through the mechanism of negative priming.

In-task priming is stronger than framing. Our experimental set-up includes two priming mechanisms: in-task priming, produced by varying the first 5 images of the task, as well as what could be called sidestream priming, in which a fictitious funder is disclosed. The intention behind this set-up was to enable a direct comparison, and our expectation was that in-task priming might produce some fraction of the effect produced by disclosing the funder. To our surprise, in-task priming produces a much stronger effect.

There is a cautionary lesson here. In-task priming, which is inherently impossible to eliminate is may be far more severe than the more overt causes of priming that more routinely attract concern during the design of an experiment. Depending on the final purpos of HPU work products, this may be quite restrictive.

Leveraging priming. In a more positive view, the results of our study suggest that the ordering of subtasks can be used as a way to direct the focus and attention to detail by HPUs. If the designer desires more coarse-grained work-products, this can be achieved by presenting successive subtasks with high diversity. On the other hand if fine-grained work products are desired, then the designer should aim to sort subtasks into "tracks" that are highly homogeneous, and assign subsens of the pool of HPUs to specifict tracks.

It is interesting to consider to what extent the output of HPUs can be driven toward nuanced detail using this technique. Consider, for example, the image labelling platform called the ESP Game. Here Two HPUs are shown the same

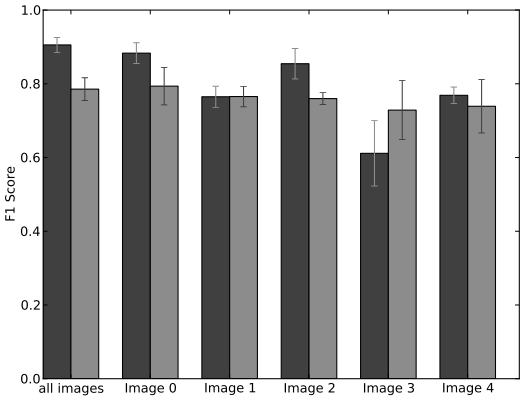

Figure 4: caption here

image and in order to derive what are in some sense the moset characteristic labels, they are asked to attempt to produce the same labels for the image. One could imagine a similar setup, but in which the designer attempts to produce the set of labels that best covers the semantic space occupied by the image, by eliciting labels at coarser and finer levels of specificity using the techniques suggested by the present work.

One interesting test of this hypothesis arises in

References

[1] Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133, 2013.

Figure 5: caption here

Figure 6: caption here

Figure 7: caption here

Figure 8: caption here