**Epistemic Annotation Guidelines 10/8/18**

**Overview**

The goal of this work is to identify and classify epistemic statements in the scientific literature and the general subject of such statements. An epistemic statement is a statement of unknowns, hypotheses, speculations, uncertainties, including statements of claims, hypotheses, questions, explanations, future opportunities, surprises, issues, or concerns (more clarifications in the task description below).

We identify these statements using epistemic cues – how these unknowns, hypotheses, speculations, and uncertainties are communicated. Generally, statements of fact or methodology are not epistemic statements because the statements tend to be certain or explain the experimental design. This is not to say that there are never epistemic statements in the methods sections (see the second abstract example). For each epistemic statement, there is also the subject of it. Here that includes statements about: function, global organization, location/context, production, structures/entities, or timing.

The task at hand has 3 parts: Given a sentence (1) identify and classify epistemic cues that signify a specific type of epistemics (2) identify and classify the subject of the specific types of epistemics (3) link the epistemic cues with its subject. The end goal of this work is to determine goals for knowledge (next steps or questions) based on the epistemic statements to provide an overview of the questions in a field and to drive literature-based discovery (finding relevant articles to the goals for knowledge).

1. **TASK AT HAND**

You will be provided with 54 numbered sentences in one document in knowtator2. I will train you in using knowtator2. As you go along please take notes in the last sheet of the excel spreadsheet provided for the epistemic type, epistemic subject, how hard the sentence was to classify, and any additional notes.

For each sentence please complete all 3 tasks at hand:

1. Identify and classify spans of text that are epistemic cues that signify a specific type of epistemics: (full definitions of types can be found in attached excel spreadsheet – Ignorance\_Taxonomy\_Information\_10.8.18 – Epistemic\_Types Sheet). Along with the definition, a knowledge goal is provided. Feel free to use this to ensure it is an epistemic statement with more information needed (next steps). Further, within knowtator2 there are synonyms and you can use the search function to help navigate within knowtator2. Also, if you think I am missing a type of epistemics please add it to the last sheet in the excel document under “Epistemic Type Notes”.
   1. **Claims/Questions/Evidence**
      1. Full unknown
      2. Explicit questions
      3. Alternative options
      4. Incomplete evidence
   2. **Links/Understanding**
      1. Possible understanding
      2. Probable understanding
      3. Superficial relationship
      4. Etiology
   3. **Future opportunities**
      1. Future work
      2. Non-urgent recommendation
      3. Future prediction
      4. Urgent call to action
   4. **Surprises/Interesting**
      1. Curious finding
      2. Unexpected observation
   5. **Issues/concerns**
      1. Difficult task
      2. Problems or complications
      3. Controversy
   6. **Questions answered by this work**
2. Identify and classify the span of text that is the subject of the type of epistemics: what is the epistemic statement about in general (full definitions of subjects can be found in attached excel spreadsheet - Ignorance\_Taxonomy\_Information\_10.8.18 – Epistemic\_Subjects Sheet). Along with the definition, a knowledge goal is provided. Feel free to use this to ensure the annotating captures that more information is needed (next steps). Each epistemic type identified in task 1 should have a corresponding subject. If you do not see one that fits, please think about a new one that would capture the subject and write a note in the last excel spreadsheet under “Epistemic Subject Notes”.
   1. **Function**
      1. Abilities
      2. Activities
      3. Roles
   2. **Global Organization**
      1. Communication
      2. Comparison
      3. Disruption
      4. Independent
      5. Pathway
      6. Regulation
      7. Reproducibility
      8. Risk
      9. Safety
      10. Model
      11. Methodology
   3. **Location/Context**
      1. Body location
      2. Environment
      3. Model/animal system
      4. Sex
   4. **Production**
      1. Establishes
      2. Mobilization
      3. Sustains
   5. **Structures/Entities**
      1. Characteristics
      2. Magnitude/amount/size
      3. Molecular nature
      4. Morphology
      5. Structure, consistency, uniformity
   6. **Timing**
      1. Age
      2. Duration
      3. End time
      4. Start time
      5. Time points
3. Link the type of epistemics (task 1) to its subject (task 2). Using the graph space functionality in knowtator2, connect the epistemic types to their subjects using the relation “is about”. I will provide detailed training of how to use knowtator2 for this task on the day of annotation. Please put all annotations from one sentence in the same graph space even if parts are disconnected. Name the graph space GS# where # is the sentence number.

**NOTES:**

* 1. There can be no, one, or more than one type of epistemics per sentence. There can also be no, one, or more than one subject per sentence. Types of epistemics and subjects can form a chain with the relation is\_about. Please also put all annotations from one sentence in the same graph space even if parts are disconnected.
  2. Each type of epistemics should either map to or be mapped to either a subject or another type of epistemics. Graphing the “is\_about” relationship helps clarify this criterion.
  3. Feel free to use the search function built in to knowtator2: make sure to read the options and be careful because terms can appear in multiple categories.

1. **2 ABSTRACT IDENTIFICATION EXAMPLES**

**Key: Yellow is epistemic type (TYPE), green is subject (SUBJECT)**

* 1. **PMC48141 abstract**

**Background**

Little is known (FULL UNKNOWN) about genetic factors (MOLECULAR NATURE) affecting intraocular pressure (IOP) in mice and other mammals (MODEL/ANIMAL SYSTEM). The purpose of this study (QUESTIONS ANSWERED BY THIS WORK) was to determine the IOPs of genetically distinct mouse strains (MODEL/ANIMAL SYSTEM), assess the effects of factors such as age (AGE), sex (SEX) and time of day (TIME POINT) on IOP in specific strain backgrounds, and to assess the effects of specific candidate gene mutations (MOLECULAR NATURE/DISRUPTION) on IOP.

**Results**

Based on over 30 studied mouse strains, average IOP ranges from approximately 10 to 20 mmHg. Gender (SEX) does not typically (CONTROVERSY) affect IOP and aging (AGE) results in an IOP decrease in some strains (MODEL/ANIMAL SYSTEM). Most (PROBABLE EXPLANATION) tested strains (MODEL/ANIMAL SYSTEM) exhibit a diurnal rhythm with IOP being the highest during the dark period of the day (TIME POINT). Homozygosity for a null allele of the carbonic anhydrase II gene (*Car2**n*) does not alter IOP while homozygosity for a mutation in the leptin receptor gene (*Lepr**db*) that causes obesity and diabetes results in increased IOP. Albino C57BL/6J mice homozygous for a tyrosinase mutation (*Tyrc*-2*J*) have higher IOPs than their pigmented counterparts.

**Conclusions**

Genetically distinct mouse strains housed in the same environment have a broad range of IOPs. These IOP differences are likely (PROBABLE EXPLANATION) due to interstrain (MODEL/ANIMAL SYSTEM) genetic differences (MOLECULAR NATURE) that create a powerful resource for studying the regulation (REGULATION) of IOP. Age, time of day, obesity and diabetes have effects on mouse IOP similar to those in humans and other species. Mutations in two of the assessed candidate genes (*Lepr* and *Tyr*) result in increased IOP. These studies demonstrate that mice are a practical and powerful experimental system to study the genetics of IOP regulation and disease processes that raise IOP to harmful levels.

* 1. **PMC5975408**

## **Background**

Prevention of mother-to-child HIV transmission (PMTCT) programs usually (INCOMPLETE EVIDENCE) test pregnant women (ANIMAL/MODEL SYSTEM) for HIV without involving (INDEPENDENT) their partners. Non-disclosure of maternal HIV status to male partners (COMMUNICATION) may (POSSIBLE UNDERSTANDING) deter (DISRUPTION) utilization of PMTCT interventions since partners play a pivotal role (ROLE) in decision-making within the home including access to and utilization of health services.

## **Methods**

Mothers attending routine 6-week and 9-month infant immunizations were enrolled at 141 maternal and child health (MCH) clinics across Kenya from June–December 2013. The current analysis was restricted to mothers with known HIV status who had a current partner. Multivariate logistic regression models adjusted for marital status, relationship length and partner attendance at antenatal care (ANC) were used to determine correlates of HIV non-disclosure among HIV-uninfected and HIV-infected mothers, separately, and to evaluate the relationship (SUPERFICIAL RELATIONSHIP) of non-disclosure with uptake of PMTCT interventions (COMMUNICATION). All analyses accounted for facility-level clustering,

## **Results**

Overall, 2522 mothers (86% of total study population) met inclusion criteria, 420 (17%) were HIV-infected. Non-disclosure of HIV results to partners was higher among HIV-infected than HIV-uninfected women (13% versus 3% respectively, p < 0.001). HIV-uninfected mothers (MODEL/ANIMAL SYSTEM) were more likely (PROBABLE UNDERSTANDING) to not disclose (COMMUNICATION) their HIV status to male partners if they were unmarried (adjusted odds ratio [aOR] = 3.79, 95% CI: 1.56–9.19, p = 0.004), had low (≤KSH 5000) income (aOR = 1.85, 95% CI: 1.00–3.14, p = 0.050), experienced intimate partner violence (aOR = 3.65, 95% CI: 1.84–7.21, p < 0.001) and if their partner did not attend ANC (aOR = 4.12, 95% CI: 1.89–8.95, p < 0.001). Among HIV-infected women (MODEL/ANIMAL SYSTEM), non-disclosure (COMMUNICATION) to male partners was less likely (PROBABLE UNDERSTANDING) if women had salaried employment (aOR = 0.42, 95%CI: 0.18–0.96, p = 0.039) and each increasing year of relationship length (DURATION) was associated (SUPERFICIAL RELATIONSHIP) with decreased likelihood (PROBABLE UNDERSTANDING) of non-disclosure (COMMUNICATION) (aOR = 0.90, 95% CI: 0.82–0.98, p = 0.015 for each year increase). HIV-infected women who did not disclose (COMMUNICATION) their HIV status to partners were less likely (PROBABLE UNDERSTANDING) to uptake CD4 testing (aOR = 0.32, 95% CI: 0.15–0.69, p = 0.004), to use antiretrovirals (ARVs) during labor (OR = 0.38, 95% CI 0.15–0.97, p = 0.042), or give their infants ARVs (OR = 0.08, 95% CI 0.02–0.31, p < 0.001) (DISRUPTION).

## **Conclusion**

HIV-infected women (MODEL/ANIMAL SYSTEM) were less likely (PROBABLE UNDERSTANDING) to disclose (COMMUNICATION) their status to partners than HIV-uninfected women (MODEL/ANIMAL SYSTEM). Non-disclosure (COMMUNICATION) was associated (SUPERFICIAL RELATIONSHIP) with lower use of PMTCT services (DISRUPTION). Facilitating maternal disclosure to male partners (COMMUNICATION) may (POSSIBLE UNDERSTANDING) enhance (SUSTAINS) PMTCT uptake.

1. **CLARIFYING CONFUSING DISTINCTIONS**
2. **Possible understanding vs. Probable understanding**

Possible understanding is a statement of a feasible explanation, relationship, or phenomenon with no claim to the correct explanation, relationship, or phenomenon. The statement is more proposing options and not staking a claim as to which is correct. This includes epistemic cues such as could, can be, have potential, it is possible, and might. In the sentence, “Small changes in the resistance to aqueous humor drainage may also contribute to diurnal differences in IOP” the researchers offer another possible contributor to diurnal differences in IOP.

Probable understanding is a statement staking a claim to the most likely explanation, relationship, or phenomenon where it has a good chance of being the case or of coming about. This includes epistemic cues such as it is becoming increasingly clear. Likely, unlikely, promising, plausible, mainly. In the sentence “It is becoming increasingly clear that many forms of glaucoma have a genetic component” the researchers stake a claim that probably many forms of glaucoma have a genetic component.

1. **Controversy vs. alternative options**

Controversy is a statement of disagreement amongst researchers; a lack of consensus; at least two possible answers are presented as results from different researchers; usually in reference to previous results and stated when results disagree with each other; contradictions. There are at least two possible answers already presented as results from different researchers. This is usually in reference to previous works findings and stated when findings disagree with each other. This includes epistemic cues such as do not agree, have been variably, some but not all, some but not other studies, have been implicated, and but. In the sentence “Some but not all human studies have reported a positive association between IOP and blood pressure” there is disagreement on the association between IOP and blood pressure.

Alternative options is a statement of multiple (at least 2) choices or actions with no stated disagreements, including statements with an implied second option such as "whether". These options have not yet been studied and so there cannot be disagreements yet about which one is correct. The statement only introduces the options. This includes epistemic cues such as alternatively, cannot be ruled out, though, either, and whether. In the sentence “An effect of anesthesia in these very old mice cannot be ruled out” the researcher indicates an anesthesia affect is an alternative option for explaining some phenomenon in very old mice.

1. **Curious finding vs. unexpected observation**

Curious finding is a statement of a surprising result, conclusion, or situation; the researchers were not expecting the result, conclusion, or situation but are intrigued by it. This includes epistemic cues such as curious, however, interestingly, surprisingly, and suspected. The researchers were not expecting the result or conclusion.

Unexpected observation is a statement of a judgement or inference from an observable phenomenon that the researchers did not expect; an element of surprise. This includes epistemic cues such as apparently, appear, appears to be, seems to be, and does not appear to be. The statement describes an observable phenomenon that they did not expect.

In the sentence “Interestingly, the IOP of strain CBA/CaJ does not appear to increase in the dark.” *Interestingly* is classified as curious finding because it indicates a result that the researchers did not expect. *Does not appear to* is classified as unexpected observation because they observed that the IOP not increasing in the dark but they are not certain if this observable phenomenon is repeatable.

1. **Non-urgent recommendation vs. urgent call to action**

Non-urgent recommendation is a statement of suggestion or a proposal as to the next best course of action, especially one put forward by an authoritative body; advice telling someone what the best thing to do is. This includes epistemic cues such as advocate, encourage, guideline, should be, require, must be, and it is important that. It is a non-urgent recommendation that does not necessarily need to be dealt with right away, there is no clear sense of urgency. For example, in the sentence, “Thus, to avoid effects of anesthesia on IOP, all measurements should be made within a window of up to 12 minutes after anesthetic administration.” The sentence recommends taking measurements up to 12 minutes after anesthetic administration, without urgency, but useful for further experimentation.

Urgent call to action is a statement calling for immediate attention including an action needed to be taken immediately or information that needs to be disseminated immediately; critical: being in or verging on a state of crisis or emergency; urgently needed; absolutely necessary. This includes epistemic cues such as call to act, critical, cause for concern, crucial, essential, high priority, imperative, needed, influential, and vital. The statement has a clear sense of urgency. For example, in the sentence, “As PMTCT Option B+ scales up with regimens that require life-long antiretroviral treatment (ART) adherence, the need for HIV-infected pregnant and breastfeeding women to disclose their status to their male partners becomes more critical in order to maintain adherence to ART.” Here *the need* and *more critical* indicate a sense of urgency that needs to be dealt with sooner rather than later.

1. **Future work vs. future prediction**

Future work is an explicit statement of extensions, including next steps, directions, opportunities, approaches, or considerations, of the described work that may be implemented at some future time point. This includes epistemic cues such as additional studies, more information… is needed, needs to be assessed, future investigation/strategies/opportunities/considerations, and candidate. For example, in the sentence “Mutations in two of the assessed candidate genes (*Lepr* and *Tyr*) result in increased IOP.” *Candidate* indicates that more work needs to be done to determine if these 2 genes actually increase IOP.

Future prediction is a statement of extrapolation of given data into the future and/or from past observations; no reference to next steps or future work. This includes epistemic cues such as if real, if confirmed, expect, supposed to, will, ultimately, would, and implication. For example, the sentence “If real, this sporadic sex difference was not dependent on age, sometimes occurring in a group of B6 mice at a particular age and sometimes not occurring in a separate group of the same age.” This sentence indicates that if the phenomenon was real, then multiple things follow (a prediction). Notice no next steps are stated.