Building Stronger Theory

In this note, I offer some general considerations to bear in mind when building a theoretical argument. In order to illustrate with an example, I use as my theory the old Latin phrase *in vino veritas* (literally, 'in wine there is truth').

- **Are the key terms in your theory clear and appropriate?**
  - Are the key concepts / constructs clearly defined and understood?
    - E.g., by wine, do we literally mean wine? Or any form of alcohol? And what about other narcotics? (this is a case where moving up the level of abstraction seems necessary)
    - And is it any amount of wine? Perhaps 'intoxication' would be a better term.
  - Are the key concepts / constructs distinguishable from each other? (If this is not the case then the claim may be tautological)
    - E.g., if we were to define intoxication as a state of uncontrolled truthfulness, then the claim 'intoxication produces truth' would be tautological. In this case, however, we can define intoxication as 'a loss of control or faculties brought on by alcohol or other narcotic substances', so the construct is distinct from truthfulness.
  - Are the key concepts / constructs appropriate for the theory?
    - E.g., In this case, it would seem that honesty may be a more appropriate construct than truth. We are going to argue that intoxicated people say what they believe to be the truth, not necessarily what is true.
  - Are the key concepts / constructs (approximately) measurable? In other words, do we have (or can we imagine) empirical variables that would proxy for them? (strictly speaking this is not necessary for a theoretical argument; but an untestable theory is also an unfalsifiable theory)
    - E.g., in this case we could reasonably proxy intoxication by blood alcohol level, and honesty by whether the person was saying things that were factually accurate
  - Is the relationship between the key concepts / constructs clear?
    - E.g. it's not really that honesty is *in* the wine, it would be more accurate to say we expect intoxication to lead to more honesty. So a better statement of our claim would be: "Intoxication makes people more honest'
    - Notice that this is a causal statement: we are not simply saying that intoxicated people are more honest (which would be an associational claim) we are saying that intoxication is a cause of their honesty
  - Is the counterfactual clear?
    - If we say 'Intoxication makes people more honest', the obvious question is: compared to what? Less intoxication? No intoxication? Some other treatment (money? threats?) So again, a better statement of our claim would be "Intoxication makes people more honest than sobriety"

        - Note that in this case we're really not predicting a linear relationship; rather we're predicting a phase shift; in other words, we're not predicting that a person's honesty will increase linearly with the amount of alcohol they consume. If we were, then we would say, "the more intoxicated a person, the more likely they are to be honest'

- **Is the argument / chain of logic leading to this conclusion valid?**
    - E.g. In this case the logical mechanism might be: Intoxication lowers inhibitions -> Lower inhibitions lead to greater honesty
  - ➢ Is the logical chain complete? Do we need additional assumptions / arguments?
    - E.g., even if lowered inhibitions make a person more honest, does that necessarily mean they'll say what they're thinking? So perhaps we need an additional argument, that lowered inhibitions also make people more garrulous.
  - ➢ Is each assumption reasonable / backed by evidence? If not, what is the basis for that assumption?
    - E.g. in this case, the assumption that intoxication lowers inhibitions seems fairly sound and is easily backed by medical evidence. But the assumption that lower inhibitions lead to greater honesty is questionable. Might not lower inhibitions lead to more lying? Can we make a logical argument for why this would not be the case?
  - ➢ Are there additional assumptions that serve as boundary conditions?
    - E.g. if a person were completely truthful to begin with, then this argument would not hold. So an important boundary condition for our theory is that the person needs to be dishonest or at least holding back some truth in the first place
  - ➢ Are the assumptions consistent with each other? Is the resulting claim an equilibrium? (note: A claim could be valid even if it was not an equilibrium, but we would need to be clear about how that disequilibrium is maintained)
    - E.g. Is the assumption that a person would get intoxicated consistent with the idea of that person having strong inhibitions? Put differently, why would someone with something to hide get intoxicated in the first place?
- **Are there alternative explanations that could produce the same conclusion and can we rule them out?**
    - E.g. It may be that intoxication lowers cognitive power, making it harder for people to come up with a lie, making them more likely to be honest by default
  - ➢ Are there additional discriminating predictions we could make to rule out these alternative explanations?
    - E.g., If the alternative theory is true, then it should mean fewer new lies, but not impact old lies; conversely, if our theory is true, it should mean dropping old lies (but potentially telling some new ones!). So we could offer a second hypothesis: "The increase in honesty resulting from intoxication is more likely to reveal old lies than reduce new lies"

        - E.g., a different discriminating prediction would be to consider a different outcome that our theory would predict but the alternative explanation would not. So we could offer a second hypothesis: "Intoxication makes people more likely to engage in inappropriate humor"
    - ➢ Are there reasons why the alternative explanation may be less plausible than our chosen theory
        - E.g., is the evidence that intoxication lowers inhibitions stronger than the evidence that it produces cognitive decline?
    - ➢ Note that we are mostly concerned with alternative explanations, i.e., different mechanisms that would produce the same result. We are less concerned with
        - Theories / explanations that would produce the opposite result, since these will work against our hypotheses (though this is only true in so far as we are testing our theory directly)
        - Other effects that our predictors might produce, or other factors (unrelated to our main construct / concept) that might produce the same outcome. In other words X -> Z or W -> Y, that doesn't really matter to us if our theory is that X->Y (so long as X and W are conceptually distinct as are Z and Y. Note that if X and W or Z and Y were correlated that would create an empirical problem, but not a conceptual one)
- ❖ **Are the implications of our theory actionable?**
    - ➢ Is the theory practically useful to its intended audience?
        - E.g., the claim "Being hugged more as a baby makes people more honest', but it's hard to see what a manager would do with that (though it may be relevant to new parents)
    - ➢ Are there implications of the theory that should be qualified?
        - E.g., it may be important to note that intoxication has many other negative side effects, so we would not want to recommend intoxication as a general means of increasing honesty