1 Memorandum of Understanding

A list of the work we've done so far, so as not to lost track of it all.

1.1 Research Questions

1. How can a computational formalism of responsibility direct the decisions made by an intelligent agent?
2. How can an intelligent agent assume the consequences of actions it makes, the decisions other agents make, and its general environment, so as to direct its interpretation of responsibility?

2 Basic Definitions and Terminology

We have some basic definitions and terminology to keep track of:

1. An *obligation* is a set of satisfaction constraints which describe something which needs to be done.
   For example: tidying a room might be constrained by the satisfaction criteria:

   ```
   {'room-tidiness': '>= 0.9', 'mess-created-elsewhere': '<0.3'}
   ```

2. A *responsibility* is an assigned and accepted obligation.
3. An actor which assigns an obligations is an *authority*.
4. An actor which receives responsibilities is a *delegee*.
5. When an obligation is assigned, information relative to the assignment is bundled with the obligation to turn it into a responsibility. This includes:

   - The assigning authority
   - A score of importance from the authority's perspective.

   Note that the score of importance is only determined when the responsibility is assigned: it might be imperative for one actor to fulfil an obligation, but not another. An example of this might be writing an essay. It might be imperative that one student write the essay, but optional for another.

6. When a delegee is assigned a responsibility, they may accept or reject that responsibility. A boolean indicating acceptance is returned to the authority.

7. When a delegee accepts an obligation, the resulting responsibility's score is interpreted into a delegee-specific score via an interpretation function (for subjectivity).

   This interpretation function can be a learning algorithm such as reinforcement learning, and the agent directs its interpretation of responsibility this way.

3 Functions to be aware of

3.1 Calculating other agents' responsibility

Calculating how responsible another agent is can be done by, for every constraint which affects a resource, sum the authority-perspective importance score multiplied by the binary form of the result of the satisfaction constraint. Divide by the number of constraints analysed.

Form a vector of this calculation for each resource analysed; we end up with a vector of the responsibility of an actor from all specific perspectives.

3.1.1 Calculating an agent's own responsibility

Perform this calculation by analysing your own consequential responsibilities.

3.2 Calculating which agent to delegate to

We don't necessarily want to just maximise the responsibility of the delegee, because a very generally responsible agent may be very irresponsible for specific resources.

We make a matrix, where one dimension is resources, and the other actions that an agent can take. We make this matrix for each agent. Each cell of the matrix is a representation of the affect of the action on the resource indicated by that cell's position.

Then, for each agent's matrix, we analyse what actions the agent can take which discharge the given obligation appropriately. For all agents which can discharge the obligation successfully, their specific responsibilities relative to the resources the obligation concerns are combined. We select the delegee by selecting the maximum specific responsibility from this pool.

How are these combined? Just summed?

3.3 Influencing an interpretation of responsibility

. . . at the moment, I think the only way of doing this that's been considered is an interpretation function; nothing concrete.