Description of relevant current research projects

My bachelor's thesis was about modelling of natural intelligence. Specifically, classical and operant conditioning, plus insight learning.

Now I'm beginning my doctorate studies and my research topic is safe artificial intelligence. Specifically, how one can develop safer behaviours by using the principle of avoiding irreversible actions.

According to current hypothesis, the goal system of AI / robot consists of two parts:

1) permissions and restrictions
2) goals / goal states

By default any change in the world is not permitted.

By giving the robot some permissions for certain kinds of changes, it can "use" – that is – cause these changes, during its "path" towards goal-state.

Permissions do not specify the preferred state of the world. Instead, they describe which aspects of the world are allowed to change as a result of robot's actions, and which not.

When robot nevertheless inadvertently changes something not permitted to change, it will try to reverse that action, if possible and when it has appropriate earlier experience.

Goals, by contrast, do have fixed specifications about which state of the world is preferred or not-preferred. Robot tries to achieve these preferred states and then optionally keep these aspects of the world from deviating in the future.

Latter means that there may exist a point in time when robot settles to rest and will not go on infinitely. That is probably also safer.

Both permissions / restrictions and goal states can be represented as setpoints for the operant learning and planning module or for any other suitable module, which behaves according to the principles of control theory.

Goals get their setpoint's preferred state from entries in a database describing the robot's goals configuration.

Restrictions get their setpoint's preferred state from sensor readings. When sensor readings change, it will be determined whether the cause of the change was due to the robot's actions, or external. If the cause was external or to the extent it was external, corresponding setpoint's preferred value will be adjusted to sensor's new reading. Otherwise, setpoint's preferred state does not change and robot tries to eliminate the discrepancy that has appeared.

Permissions are essentially disabled restrictions.

In order to decide about above mentioned source of change, one has to have a good causal understanding of one's action's results and also understanding about external causes. That problem involves learning – learning correct causal relations, but not co-occurences or events that follow each other but are not causally related. One needs to assign credit appropriately.

Motivation for participation

My motivation for participation is finding the best possible and best practical credit assignment method. I have a hypothesis, that in natural thinking, in classical conditioning the aspect called “blocking” deals with this problem better than other learning methods. In order to verify this hypothesis experimentally, I need to understand and apply various contemporary and good learning algorithms, then compare their results to the “blocking” method.