A selection of "big" research questions I have:

Hierarchical agency, horizontal negotiations, and social constructs

Agency exists at multiple levels of abstraction. Within each—using all terms loosely—negotiations between agents produce intelligent action:

- Inputs to neurons negotiate to produce activations
- Groups of neurons negotiate to produce subconscious thought
- Subconscious thoughts negotiate to produce natural language thought
- Competing subconscious and natural language thoughts negotiate to produce human action
- Groups of humans negotiate to produce the behaviors of economic entities (firms, governments)
- Economic entities negotiate to produce global-scale action (m&a, national action, int'l trade)

Although we typically characterize intelligence to lie in second to fourth bullets, one might argue that firms and nations exhibit "economic intelligence". Can we create a theory of intelligence that captures the different types of possible negotiation? Can we draw inspiration from what economics and game theory tell us about the latter three bullets to inform our models of the former three? (cf. GANs) How do multiple levels of negotiation interact, and how can design be used to optimize this interaction?

For instance, laws and governments are examples of artificial constructs created to govern certain aspects of the negotiations involved in the final 3 bullets. Can such constructs be pushed down the hierarchy (e.g., to metacognition)? We likely don't want to enforce thought crime for humans (even if it were possible), but what about for robots?

Artificial ethics; balancing social vs individual utility

Ethics are closely tied to intelligence because they are an essential part of the aforementioned negotiation: in order to negotiate, we must develop certain expectations about the other parties, which depends heavily on their social behavior (cf. foe/friend-Q and correlated Q-learning). How can we design artificial agents that behave ethically?

To my knowledge, how to best model artificial utility is an open question. Should everything be left to individual utilities, as it necessarily is for humans, or does it make sense to endow artificial agents with constructs of social utility, since we can? How can we design those constructs to be socially efficient? What definition of social efficiency should we use?

Modeling cause and effect; credit assignment; justification; logical reasoning

Explicit sequential reasoning is at the core of intelligence. We would like to grant our agents the ability to make forward-looking hypothetical statements like "if this then that" and backward-looking causal statements like "this was caused by that". Without this ability our agents will not be able to justify their decisions, and will have limited faculties for logical reasoning. How can we address this? This is a hard problem, to which a general solution may imply artificial general intelligence.

The first step, however, is clear; we need a way to model cause and effect. One immediate problem is that causation is not well defined in a real world (uncountable/continuous) setting: in legal reasoning, there are more than one type of cause, each with multiple possible criteria. The concepts of *source traces* (my contribution) and *eligibility traces* work together to endow reinforcement learning agents in discrete MRPs with a model of "actual causes", both immediate (eligibility traces) and potential (source traces). Can source traces be extended to control? (Likely yes.) To approximate state spaces? (Likely yes.) Can these concepts be leveraged into a model of proximate causation? (Possibly.) Into a system for explicit justification? (Possibly.) Can we generate an inverse model of effects that, like eligibility and source traces, is invariant to the choice of time scale? (Yes; see successor representations.) Can we combine it with generative models to grant agents the ability to generate hypothetical states? Can that ability be bootstrapped into reasoning?

One can look from causes to effects, or from effects to causes. Closely related is the concept of Bayesian statistics and the relationship between predictive and generative models. Humans have a natural ability to invert causes and effects (do Bayesian inference) at different levels of abstraction (over arbitrary sets), which is accompanied by automatic generalization to the inverse and the ability to reason simultaneously over partially known causes and effects. Can we design a neural architecture that does this as a byproduct of prediction or generation, or both, and at different levels of abstraction?

Link between memory and imagination/generation

It is true that imagination is a predictive model based on experiences (i.e., memories), but the link between imagination and memory runs deeper, in the following sense: for purposes of logical reasoning and thought, memories and generated experiences are (near) interchangeable. If we want to describe what a bear looks like, we can retrieve a memory of a bear, or we can imagine a bear in our head. How are the two different? How can we design agents that use memories and generative models interchangeably? How should they choose which to rely on? Importantly, since they *learn from memories* (cf. experience replay), is it possible to *learn from generated experiences* (cf. Dyna), and can we bootstrap this into out-of-experience learning? E.g., can we design an agent that will use generated experience to prepare itself to solve a described problem that it has not yet experienced? Cf. transfer / zero-shot learning.

Locality of learning and function approximation

Human learning is extremely local: if we learn how to perform task A, it usually does not affect our performance on task B (unless of course, by improving it via generalization). Is there a way to characterize the “locality” of a learning algorithm? (cf. Gordon 1995) To define and measure it mathematically? How can we build artificial agents that learn locally and can adapt to multiple tasks without suffering from catastrophic forgetting? (Elastic weight consolidation is a start, but clearly not sufficient on its own - Kirkpatrick et al. 2017)