MAPLE

THE HARD PROBLEMS OF AI ALIGNMENT
DRAFT, OCTOBER 2022

This essay is about two hard problems that seem to lie at the center of the AI alignment puzzle. These two problems seem to have a different character of difficulty to other problems. Very roughly, they are:

1. Where in any system are terminal values encoded?
2. What is agency, and where in any system can it be found?

These problems seem to be hard in a way that is different from the way that designing a parachute, or launching a satellite, or characterizing the runtime performance of a web

server is hard. The hard problems of AI alignment seem to defy the basic frames in which they are defined, and yet also seem to be on the critical path to solving an urgent problem.

The hard problem of measuring value	The hard problem of identifying agency
Where in any system are terminal values encoded?	Where in any system can agency be found?
How can we build any information processing system that acts in service of what is good?	How can we build any information processing system that is agentic?
How can our terminal values be encoded into an algorithm?	How can an algorithm be alive?
How can we build an algorithm that adopts our values over time?	How can we build an algorithm that retains goals over time?
What kind of object becomes more and more good over time?	What kind of object becomes more and more powerful over time?
How can any machine be good?	How can any machine be agentic?
What is good?	What is agency?

This essay is organized as follows. Sections one and two are about the hard problems and why they are, in fact, hard. Section three is about hard problems in general. Section four is about how to work fruitfully on hard problems. Section five is about ways forward from here.

The hard problem of measuring value

Where in any system are terminal values encoded?

It seems that any artificial intelligence system would contain an algorithm for choosing actions. But how could any algorithm encode what is good within its decision criteria? We might align AI systems with our proximate goals and use human oversight to update their goals over time, or we might encode all that we value into a single AI system, or we might

devise a framework for cooperation among many AI systems and many humans, or we might devise an AI system that incorporates end-to-end models of human behavior in specific ways. All of these approaches run into the basic problem that we don't have a trustworthy formulation of what it means to take ethical actions.

In classical AI safety, this problem was cast as value learning. Humans were modeled as agents with terminal values, and the value-measurement problem was to find these terminal values and build an AI system that pursues them. But modeling humans as agents with values is just one approach to the general problem of constructing an algorithm that does what is good. More recent work in AI alignment has ventured outside this paradigm, and yet the value measurement problem shows up in these approaches, too.

Approaches within "indirect normativity" build an end-to-end model of human behavior and attempt to solve the value measurement problem without ever eliciting "terminal values" from it. This includes iterative distillation and amplification, AI safety via debate, factored cognition, and so on. Of these approaches it can be asked: what is the functional relationship between the overall system and the model of human behavior, and why is that particular relationship good? For example, humans themselves can be unclear or dishonest about their goals, and their goals may not be coincident with what is good, so it is assumed that models of human behavior may also be unclear or dishonest about their goals, and hold goals that are not coincident with what is good. Indirect normativity addresses this by organizing models of human behavior into computation graphs alongside learning and external oversight. But the central question that always seems to remain is whether any particular computation graph is aligned with what is good, or with our intentions, or with our terminal values. It may be that asking this question in terms of "what is good" is very different from asking it in terms of "intentions" or "human values", and yet in all three cases – and for other formulations – we seem to have no handholds at all with which to assess such an alignment question. This is the value measurement problem itself, showing up within an attempt (indirect normativity) to sidestep the value measurement problem.

Let's consider a second way to move beyond the classical AI safety formulation of value learning. Under interaction games we give an AI the goal of discovering what it ought to pursue using interactions with humans. The AI is designed to maintain uncertainty at all times about its own goals, and to seek evidence about its own goals. That evidence comes from interactions with humans. But how actually does an AI update its beliefs about its true goals as it interacts with humans? Any particular method for doing this requires some prior assumptions about the relationship between what is observed by the AI – sensor data, human answers to questions, etc – and the quantity being inferred – goals. Within these prior assumptions there are base-level modeling assumptions such as viewing interactions with humans as Markov decision processes, as well as quantitative priors needed to separate values from other aspects of cognition. But through deliberate choice of priors, one can infer most any values from most any data. Normally we turn to simplicity priors but is that really appropriate when inferring values? By what method would we decide whether simplicity priors are appropriate for values? In general, what method ought we give to an AI to decide what is good? Here, again, the value measurement problem shows up within an attempt (interaction games) to sidestep the value measurement problem.

This extremely difficult question – how to assess, in a systematic way, what is good – seems to re-arise right within every attempt to answer it. This essay does not propose an answer; instead we seek to clarify what is difficult about this problem, and the practical pitfalls that arise when working on it.

When we ask the question of the value-measurement problem, we are using a frame to view the world. One such frame is: to whatever extent there is something worth aligning artificial intelligence systems with, that thing is encoded within physical humans. The question asked within that frame is: how can we encode the same thing into an artificial intelligence? This question, we claim, is both reasonable and urgent. It is reasonable because, if what is good can be encoded into humans, then why would it not be possible to encode it also into a machine? It is urgent because the world is rushing forward with artificial intelligence systems that are not aligned with anything much at all, and we really

do need a better way forward. Any kind of philosophical punting of the question runs immediately into the practical urgency of the situation: what exactly can we tell the world's AI engineers when they ask us how their systems can be turned in a beneficial direction?

Why the problem is hard

The value measurement problem seems to be difficult in a way that is quite different from other difficult problems in artificial intelligence.

Consider the problem of building an effective planning algorithm for an autonomous car. This is actually extremely difficult. One needs to consider not just the road situation as it exists now, and not just the road situation as it is likely to exist over the next few seconds, but also the way in which the actions of the autonomous car itself will affect the behavior of other road vehicles. Planning, therefore, cannot be separated from prediction. The algorithmic challenge of searching a joint planning/prediction space within the time constraints of live autonomous driving is high. But there is an engineering process by which we can solve this problem. We can build and test algorithms on real cars and in simulated environments. We know how to measure success. In contrast, if we knew how to measure success in the value measurement problem then we would already have solved the problem.

Consider a different example: the problem of tracking the movements of a camera through space, using only the video stream from the camera itself. This, too, is a difficult problem. The only way to measure the camera's motion is to estimate it from the movement of stationary objects in the camera's field of view, but cameras cannot directly perceive distance to the objects that they are viewing, so one must use two or more frames to estimate it, which requires already knowing the motion of the camera between the time at which the two frames were captured. There is a way to solve this problem using Bayesian estimation, but it is a high-dimensional problem over a non-Euclidean search space, and it is difficult to implement a robust solution. Nevertheless, it is possible to collect data and assess the effectiveness of any particular implementation by comparing it to ground truth

data. In the value measurement problem, if we merely asked humans to assess situations that arise within a dataset, we would immediately face questions such as: How should the questions be worded? Which people should we ask? Should we ask how much they like a certain situation, or to assess a situation ethically? How much time should we give them? It is not that these questions are impossible to answer, it is that we are lacking any basis for such important choices, and finding this basis contains the whole of the value measurement problem all over again.

One further example: imagine building an AI system that decides which scheduled meetings should be held in which conference rooms, and suppose that there are so many consequences of scheduling meetings in different rooms – walking time, room size, privacy, video conferencing, security, lighting – that we do not expect to write down a complete objective function at the outset, or perhaps ever. Still, we can begin with something and look in a practical way at the consequences. We can start with a first hash of the objective function, put it into practice, and then ask the meeting participants to tell us what happened. We can get a sense for what factors are most important to add to our objective function. Over time, we can probably do a pretty good job of designing this AI system. This is because we know how to ask a person about their experience with scheduled meetings, and we know how to set up an iterative engineering process that does a pretty good job at solving this problem. On the other hand, if we set up an iterative engineering process for the value measurement problem – say by building AI systems with the intention of doing good and then assessing regularly whether and in what ways they are doing harm – we would have to ask whether that process is effective at solving the problem. There are many iterative-improvement processes in the world that have led, historically, to enormous harm, and some that have led to enormous good. How can we ensure that our particular iterative-improvement process will lead to a correctly aligned artificial intelligence? This question again contains the whole of the value measurement problem, because to assess any process we would require something to measure it against, and the thing we would measure it against would be some measure of what is good.

I once built a computer game that simulated a bouncing ball. I didn't exactly have an objective – formal or informal – in my mind. But I could run the program and see what it was doing, and I followed an internal sense of fun. The result was not a solution to the value alignment problem but a simple game. If we use some internal sense to guide our actions, we better be sure that we are using the right internal sense, or else the result might turn out similarly trivial. What basis do we have to suppose that one or another internal sense is the right thing to follow? It is not good to leave this kind of question unanswered, and answering it would seem to include the whole of the value measurement problem.

In summary: while engineering problems are typically solved with respect to something measurable, the value measurement problem asks: what is the right thing to measure when designing powerful machines? In this way, the difficulty of the value measurement problem seems to be of a different character from problems that we already know how to measure.

The hard problem of identifying agency

What is agency, and where in any system can it be found?

We would very much like to understand how it is that any entity can exert such great influence over the future as humans appear to do on Earth at present. How can we recognize the particular kind of activity that exerts such influence over the future? All activities influence the future in some way, but humans at present seem to exert influence in a way that is different – if only quantitatively – from the influence exerted by a video recorder or a robot vacuum or an air conditioner. We would very much like to understand this kind of activity in sufficient detail to build machines that exert influence in a precise, efficient, and predictable way. Beyond that, we would like to detect the capacity of any machine to exert great influence over the future so that we can avoid unwittingly deploying high-influence machines into the world.

In classical robotics literature, agency is divided into perception, planning, and control. But a robot vacuum does perception, planning, and control, just as (presumably) a powerful AI would. What characterizes the difference in scope-of-action of these two systems? Given a blueprint for an AI system, how would we calculate its scope-of-action on this scale?

Discourse in AI alignment often connects agency with expected utility maximization. But how – even in principle – would you determine whether a neural network found through reinforcement learning was doing expected utility maximization? More precisely, how would you place such a neural network on a spectrum from rock to robot vacuum to powerful general intelligence? What even are we talking about when we refer to such a spectrum? It is not computing power alone, because an inert GPU has a lot of computing power but is not very agentic. It is not information alone, because a USB drive containing a copy of wikipedia has a lot of information, but is not very agentic. We are not arguing here about definitions of intelligence or whether intelligence exists, we are asking how to identify the capacity of any machine to influence the future. If IQ tests are the right way to characterize agency, then how actually do you apply an IQ test to a neural network? It cannot be done by simply inputting IQ questions to a neural network because we wouldn't know whether the neural network was optimized to answer IQ-type questions in a deceptive way.

Under Cartesian assumptions, AIXI and the Universal Intelligence Measure do a pretty good job of characterizing agency. A Cartesian assumption means a formal separation between agent and environment, with a well defined information channel between the two. But the situation changes drastically when there is no such separation, as Scott and Abram described beautifully in their work on embedded agency.

Under the dynamical systems view of optimization, agency is the kind of object that, when inserted into an environment, turns it into an optimizing system. But what kind of object does that? This is more like a restatement of the original question than an answer.

Why the problem is hard

Different from the hard problem of measuring value, the difficulty with agency is not that we are fundamentally lacking a basis for knowing what to look for, but that every place we look seems not to contain agency. This seems to have a really different character from other kinds of characterization problems. We will work through a few examples of difficult-but-not-hard problems and contrast them to the problem of identifying agency.

Consider the problem of determining whether an object is hot, with the goal of picking up the object. We can place a thermometer next to the object, and measure whether it is hot. It might be that the object is cool on the inside, but that does not really matter for our purposes because we are only going to touch it on the outside when we pick it up, so the outside is also where we should hold the thermometer. It also does not matter much whether the object internally consists of solid rock or intricate clockwork – we only really care about the interface, which in this example is the physical surface. In contrast, to discern whether an object is agentic, we would need to examine the internal workings of the object.

Now consider the problem of characterizing the runtime performance of a web-server. We might run the web-server and observe the time it takes to respond to web requests. In order to do this we may need to execute every line of the web server's code. Therefore, we need to understand enough about every line of code to execute it. In a certain way we need to "touch" every part of the program, unlike the heat example where we only needed to touch the physical surface. But we do not need to understand every design decision and every architectural aspect of the web server's code. Understanding the low-level semantics of each line of code is a kind of "abstraction layer" that lets us assess the runtime performance of the web-server without understanding everything about it. In contrast, there seems to be no closure to the understanding we would need to assess agency. If we merely test a computer program in a series of simulated environments then we have no way to know whether its behavior in that finite set of test cases is representative of how it

will behave in general, and it is precisely the quality of agency that correlates with computer programs that will behave differently in test versus real-world environments.

Unlike seemingly any other quality, the quality of agency cuts off any possibility of black-box assessment. Any time we try to assess agency in terms of a black box, we immediately run into the question of whether there is an agent there that is deceiving us, and in order to resolve that we are always turned back towards a gears-level understanding of the object.

We are not saying that there is an agent hiding in some microscopic computer chip hidden within some object. We are merely assessing whether the machinery that we can see constitutes an agent. Normally, even when we can see the machinery, we still apply some kind of abstraction layer to assess qualities of the thing, such as holding up a thermometer or executing a computer program. It is precisely the quality of agency that seems to defy any such use of abstraction layers.

We are not asking whether an object is conscious; we are only asking whether an object contains the potential to greatly influence the future, in the way that humans seem to, and digital alarm clocks seem not to. It is precisely the kind of intelligence that seems to be steering the future that also seems to have the potential to hide from – seemingly – any conceivable black-box testing apparatus. This seems to distinguish the agency identification problem from other characterization problems.

Suppose I wish to assess the amount of waste produced by a nail factory. To look at the factory as a “black box” would be to measure the number of nails coming out of the factory in comparison to the amount of waste coming out. But if the factory staff were aware of being measured in this way then they might retain some waste internally, invalidating the external waste measure. A more reliable way (perhaps) would be to go into the factory and

measure the waste produced by each of the factory's processes. If there is any concern that these processes might themselves be concealing waste then you could "go inside" those processes and measure waste at a still more granular level. In the case of measuring waste, there is a thing to measure at each of the terminal nodes (waste), and a clear way to combine low-level measurements into a high-level measurement. In the case of agency, what is the corresponding methodology? This is the agency identification problem.

Even logical proof may be a kind of black box behind which agency can lurk. If we have some logical predicate p that we take to indicate that an algorithm is benign, then an agent need only set itself up to take actions that cause p to be true conditional on the existence of a proof for p in the formal system that we are using to do our own reasoning, and Löb's theorem tells us that we will prove p[1]. Among all objects we might encounter in the world, it is precisely agents that can do this, because the Löbstacle applies precisely when we are proving things about the behavior of objects that condition their behavior on the output of algorithms (and in particular when those algorithms search for proofs in the same logical systems that we use to examine them).

Hard problems in general

Hard problems are:

1. compelling within the frame in which they are posed; and
2. problems for which there is no way forward that we know of.

Compelling means that the more we look at the problem, the more the problem makes sense. A problem that is not compelling can be – and often should be – cast aside.

Compelling also means that the problem has practical importance. Theoretical problems can have practical importance, as the hard problems of measuring value and identifying agency do.

[1] Is it always possible for an agent to set itself up this way? We're not sure.

Most importantly, hard problems are real. They are not mere paradoxes. Their solution does not consist of dissolving some question. They demand a practical way forward. When we have a real problem, we can assess whether we have a real solution by checking whether we still have the problem after applying the solution. In this way we can stay connected to something real despite enormous difficulty.

When a problem is hard, it seems not possible to solve by the current tools we have available, in the current way we know how to use them. We do not even seem to have a grasp on how to apply the tools we have to the problem. It seems untouchable in some way, with no incremental understanding gained from looking at it in the frames we have available.

We suspect that for these hard problems, we can reasonably ask them in our current frames, but perhaps they cannot be answered within those frames. It seems plausible that we will have to leave our frames behind and find new frames with which to answer the problem, potentially entailing a paradigm shift. Yet we cannot know if this is the case either, as we have no insight on any way forward for these problems, except that it doesn't seem to be anywhere we've looked yet.

The value measurement question is sometimes asked in the following frame:

> To whatever extent there is something worth aligning AI systems with, that thing is encoded within the physical world.

Answering the question may require abandoning that frame.

Related concepts

In this section we compare the concept of “hard problem” to paradigm shifts, paradoxes, and koans.

Paradigm shifts

It seems to us that studying hard problems can lead, in certain cases, to the kind of paradigm shifts described by Thomas Kuhn. Perhaps there are other ways that paradigm shifts can begin, or perhaps all paradigm shifts originate with an important practical problem formulated in a previous paradigm but unsolvable in that paradigm. What we are most interested in is the means by which hard problems can be solved, and the pitfalls that distract us from their solution. Kuhn’s framework, in contrast, mostly describes the social phenomenon of paradigms and paradigm shifts, which are the consequence, perhaps, of hard problems.

Not all hard problems are technical. It is possible to face a hard problem concerning how to proceed with one’s life in a very personal way. The insights provoked by solving such a problem might be extremely practical, looking nothing at all like a scientific revolution or social paradigm shift, and yet be profound.

Paradoxes

The word “paradox” can mean simply “contradiction” – a straightforward proof that one’s starting assumptions are incorrect. Such a “paradox” might be surprising, but it need not represent a problem at all, much less a hard problem.

A “paradox” can also be a seemingly absurd conclusion drawn from seemingly reasonable starting assumptions. Such a paradox forces one to either accept the conclusion, reject one of the premises, or reject logic itself. Such paradoxes can guide investigation of certain areas of philosophy and mathematics, but many such paradoxes lack a compelling practical aspect. The various (apparent) paradoxes concerning infinite sequences, for example, are instructive teaching tools, and do in fact have resolutions, but one may not be driven to solve them by a real-world need.

Theoretical problems can be hard problems. The hard problems described in this essay are theoretical, yet they are connected to a practical problem – that of resolving the precarious world situation around AI. The practical aspect is not the only reason that they are worth solving, but if a problem lacks a practical aspect then it seems to fall apart when severe practical challenges stand in the way of a solution. Paradoxes may or may not have a practical aspect, whereas hard problems always have a practical aspect.

Koans

Koans are a teaching method developed within Zen Buddhism. Hard problems are not part of any teaching method; they are problems whose solution we seek for their own reasons.

Pitfalls

This section explores common pitfalls when working with hard problems.

Deny that there is a problem

Some people say it doesn't matter whether the machines we build are aligned with our ethics. It is said that our machines are our offspring, and will be incomprehensible to us, and any attempt we make to shape their ethics will be as misguided as a forager tribe ten thousand years ago attempting to shape the ethics of the coming agricultural-industrial world. Therefore, it is said, we need not solve the value measurement problem in order to proceed.

Others say that there is a universal tendency towards greater and greater complexity, and that complexity is what is ultimately good. Therefore, they say, artificial intelligence will ultimately be good, and the value measurement problem is not a central impediment to anything.

On agency identification, some say that there is no general structure of agency, and hence no useful theory to be found. The quest for a general understanding of agency is said to be like searching for a general theory by which any company can become more productive, which is taken to be something about which no general understanding can be found.

In contrast, we have the sense that the value measurement and agency identification problem are practical impediments to resolving existential risk. It is unethical to build a powerful artificial intelligence system without any basis for believing that its consequences will be good. If our species is going to build powerful artificial intelligence systems at all, it seems that to avoid existential risk we must know what values those systems hold and whether those are good values, and the scope of the systems' power to affect the world.

Stop paying attention to direct experience

It seems that hard problems set up a kind of impasse between our direct experience and our understanding of the world through a particular frame or model.

In the case of the value measurement problem, our direct experience tells us that there is that which is worth protecting in the world. A commonly held frame is that to the extent that there is anything worth protecting, it must be somehow encoded within humans. This sets up a kind of impasse as we go looking for the physical encoding of the thing that is worth protecting (or for an algorithm that will find it) and are unable to locate it.

In the case of the agency identification problem, our direct experience tells us that there is that which has the capacity to influence the future. A commonly held frame is that the only thing that affects anything is the physical world. This again sets up a kind of impasse as we go looking for the basic structure of the thing that has the capacity to influence the future within physics, and are unable to find it.

One way out of these impasses is to trust one's model so completely that one stops paying attention to direct experience, and, because it's not explicitly modeled, believe there isn't anything worth protecting, or that there isn't anything anywhere with the capacity to influence the future.

We can take, for example, a nihilistic ethical stance and flatly deny that anything anywhere is truly worth protecting, or that there is nothing anywhere that has any real agency. The only way out of nihilism is to pay some attention again to our direct experience.

Stop paying attention to frames

If we don't examine our frames carefully, it is easy to imagine that everything actually checks out, even when it doesn't. The AI alignment problem forces us to get really precise about how we think things work, because we want to make engineering-level proposals about how to build intelligent systems that are beneficial to all. In the absence of such an urgent and practical task, it is easy to be vague about certain small but crucial details of our understanding, and as long as those crucial details remain vague, the rest of the conceptual apparatus can rest unchallenged.

Agency in a system is sometimes modeled as coming down to perception, planning, and control. Having such a model, we may think we will be able to determine agency in any system. But this is not a sufficient answer to the hard problem of agency. We do not know a way to determine, even in principle, whether a neural network produced by machine learning does perception, planning, and control. This is not an infinite sequence of "but what really is that?" questions – we actually want to know how to identify agency in any system. If we do not pay attention to our frames then it really seems as if our frames are sufficient for the task.

Conclude that the hard problems are impossible

After looking for a long time without success, it is tempting to give up. But we should not give up. It is said that it is better to die than to live defeated, and it does seem that we have the opportunity to simply not give up.

There is a subtle kind of giving up in which we conclude that the hard problems are impossible, and go looking for some different kind of work to do, still working nominally on our original goals, but no longer addressing the hard problems themselves. It may be quite reasonable to find different kinds of work to do, but it does not make sense to justify such a decision on the basis that the hard problems are impossible, because the hard problems are not impossible.

We know that the hard problems are not impossible because we ourselves have both the capacity to influence the future and the capacity to protect that which is worth protecting. If there is some reason that these capacities can never be put into machines then we should understand that, and such an insight will itself greatly clarify the alignment problem.

Take a non-solution as a solution

Ultimately, we either avert existential risk due to AI, or we do not. If we have a solution to the value measurement problem, then we should be in a position to use it if and when the need arises. If we have a solution to the agency identification problem, then we should be in a position to say something about whether particular machine learning models are agentic or not.

I came across a paper recently about a category theoretic formulation of agency as a property of dynamical systems. The abstract was very exciting to me as it suggested a full solution to the agency identification problem. As I examined the paper, I was impressed by the conceptual scaffolding that was put in place, and the elegance of the formalisms, but as I tried to work out how I would apply it, even in principle, to the problem of identifying agency in machine learning systems, I found that the theory being presented would almost

certainly say “there is a possible interpretation of this system as an agent” for any neural network of any non-trivial complexity. The material in the paper is beautiful and worthwhile, but it is definitely not a solution to the agency identification problem yet.

Give up on finding a real solution and simultaneously take a non-solution as a solution

It is possible to simultaneously deny that a solution will ever be found, and simultaneously take up a line of work – usually one that accords with one’s society – that will at best be ineffective and at worst deepen the problem.

This pitfall seems to play out when researchers decide that the entire AI alignment community is misguided in some crucial way, but then, rather than looking for a better way forward, they fall back on our society’s standard approach, seemingly deciding that it is more viable than they had originally thought. Now of course, if one encounters evidence that the standard approach is viable then one ought to update promptly, but how likely is it that this would happen just at the moment that one decided that the niche approach is never going to work? Taking the standard approach as a default has a consequence and does not seem to give us a solution to the hard problems.

Become lonely

It is difficult to go far in a direction different from that of one’s society. Working on hard problems seems to lead in this direction. In the single-minded search for a solution, one exits frames that one’s friends and family take as fundamental. If connection is about being with people who have similar perspectives to oneself, then working on hard problems seems to reduce connection.

There are many things to say about this. First, loneliness does not seem to be an illusion. Humans really do connect, at least in part, on the basis of shared perspectives, and at their most helpful, friendship circles can be incredibly helpful.

Second, connection based on shared perspectives is not the only way to connect. We can in fact find nourishing connections with strangers who don't speak our language, with non-human animals, and even with mountains, rivers, and rocks.

The basic pitfall with loneliness is not that the experience of loneliness is antithetical to working on hard problems, but that the cost of loneliness can seem high, and leave one vulnerable to exploitation by those who recognize and take advantage of it in others.

Focus only on measurable progress

Unfortunately, the frames we are trying to get out of seem to define what measurable progress ought to look like in such a way that, if we do the kind of work that leads to measurable progress by the standards of the frame, we stay within the frame.

Consider a person who is trying to rid themselves of a religious doctrine, but who always does the kind of work that is measured as progress by the standards of that religious doctrine. Or consider a person in an abusive relationship who has the goal of getting out of that abusive relationship, but who measures progress towards that goal according to whether their abusive partner approves of their "progress". These people, despite holding the goal of getting out of the frames they are stuck in, and despite working hard towards that goal, may nevertheless fail to get out of the frames they are stuck in.

Part of the frame of the value measurement problem is the view that what is most worth protecting is encoded into the human brain. A common, though not logically necessary, corollary is that we access that which is most worth protecting through our experience of wanting things. A further common, though again not logically necessary, corollary is that if we're not getting something that we want, then we're not truly making progress. But if we measure progress by whether we are getting something that we want then we will avoid experiences like deep confusion, which are often (though not always) something that we really do not want. But this whole edifice is a feature of the frame that we are working

within – the frame that says that what is most worth protecting in the world is something that is encoded into humans. What if getting what we want is not a good measure of progress?

Believe that a solution has already been found elsewhere

All sorts of research groups and companies claim – often in a roundabout way – to have some way forward on the hard problems. Blurbs on websites sometimes give the impression that the hard problems are not just solved but were solved long ago, and the focus of research has moved way past them. It's difficult, then, to work on these hard problems, when it seems as if we're three steps behind the research forefront.

I have the sense that someone somewhere really might have a complete solution to the value measurements and agency identification problems. It's not impossible. Perhaps a solution was found by someone yesterday and the whole issue of existential risk due to artificial intelligence will be resolved by the end of this year. We cannot definitively rule it out.

There is, however, a big incentive for companies and research institutes to signal that they not only have moved past these questions but have moved way past them, and this incentive puts us in a world where everyone has a solution to everything, and yet nothing really gets solved. A true solution to these hard problems ought to resolve some really significant issues in the world. The best guide to whether the problems have been solved may be whether those issues are still present.

Perpetual waiting

It is a common pitfall when working on anything big and hard to avoid working directly on the problem by various means of waiting. To wait for the right time, circumstance, conditions to be in place, setting up ever-shifting targets of when to start, that never get met. To wait by always working on some easier task that seems related and maybe important, but ultimately distracts from the main issue, and there are always endless tasks

of that type being generated. There is a kind of wishful thinking, perhaps, that during this period of avoidance, work towards the true hard problem will somehow be done. But hard problems do not get solved by waiting. Some problems in life get solved by waiting because they become obsolete due to changing conditions. These hard problems do not seem to work that way. In order to make progress on these problems, some direct effort must be applied, at the very least, to be holding the hard problems as the true task at hand, without avoidance.

Believe that the hard problems will be easily explained by the current method of research

If you believe that everything, including values and agency, ultimately is simply reducible to physical processes, then it is tempting to think we are close to solving these hard problems by way of understanding the brain's structure. Although neuroscientists would still currently say the brain is mysterious and we are far from understanding how it works completely, if you believe that understanding will come simply from knowledge of the physical structures of the brain and the rules of chemistry and physics, it is possible to imagine that soon we will have that all mapped out, and things such as values and agency will be easily located and understood for use in AI. It is also possible to imagine that though we may not be able to understand how the brain gives rise to values and agency, computer models using our knowledge of the brain's structure will be able to correctly model values and agency without us understanding their nature directly.

But what if things such as values and agency are not to be found or understood simply through physical processes? We do not have proof of that and making such an assumption may be inhibiting us from really working to get answers to the hard problems which would enable the best outcomes of our development of AI.

Moving forward

Staying with what matters

So what is the way forward to work on the hard problems of AI that avoids these pitfalls and truly get answers we can use to make the best of the future of AI? The pitfalls section describes the many ways researchers typically fall off target from working on the hard problems. The antidote is to stay focused on what matters. This may sound simple but there is a deep practice to the art of staying with what matters.

We must have a strong ethical foundation to know what matters in the first place. We cannot presume to know what is good for the world if we are not in the practice of doing what is good in our own lives. Clarity of seeing is essential. Without ethics and clarity, we can get confused about what is worth doing. We may pick rather arbitrary goals, or goals that prioritize something else, like progress, ease, fun, money. To see what matters with clarity, we must have equanimity, the ability to stay attuned to the current situation without shying away out of discomfort and without grasping for something else we want. With patience and ease, we must face the problem and really look at it to see what is there. We must be able to stay focused on this task, for which concentration is essential. We must also have a very broad awareness of the conditions of the situation, as fully as we can, so that we have an accurate context to work with.

How do we cultivate these skills of clarity, equanimity, and concentration and use them for the ability to stay with what matters? Meditation practice offers systematic techniques for developing these skills of mindfulness. The quality of being able to stay with what is relevant and attend to that in appropriate ways is absolutely learnable, yet is not the state most of us would find ourselves in at any given time. With practice we can cultivate these skills to a degree unknown by most people, and then apply them to this work. In fact, it is this work on hard problems which truly matter that provides an appropriate motivation for such practice. The two go very well together, and yet, most researchers are not intentionally bringing mindfulness to bear on their research.

Breaking out of our frames

Einstein said, “If I had an hour to solve a problem and my life depended on the solution, I would spend the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes.” Einstein was known for asking questions and how essential this was to his creative thinking. The question we ask often is coupled to the frame we are in.

As mentioned in the section on ‘what are hard problems’, to do this work we may need to break out of the frames with which we initially view these problems. It may be more than simply new knowledge we need, but rather, a new paradigm, even a new cosmology. How does one see outside the frames one is in? Does it happen by chance, by accident or a stroke of luck? Is there a systematic way to do it or to increase the likelihood it happens?

While some frames we hold are known by us, many are hidden aspects of our implicit worldview. The frame may be invisible to us as much as the water a fish swims in, taken for granted as the way things are. When a frame like this is broken out of, the frame becomes clear in retrospect. What was once a hidden assumption becomes an object to behold among objects, clearly seen as a particular view. These are the frames we think must be seen and broken out of to make progress on these hard problems.

Seeing the frame clearly is a requirement of being free of the frame. To see one’s frames clearly, it can be helpful to interact with people with other frames, paying particular attention to points of disagreement and especially confusion. Paying particular attention to disagreement and confusion within one’s own views is another way. Rather than sweep these under the rug or dismiss them, turning directly towards them and applying the tools of mindfulness can yield great insight. Cultivating a stance of non-attachment, open questioning of one’s views is very helpful as well. If we are afraid to ask ‘dumb’ questions, we may never ask the simple questions Einstein asked which led him to his deeper discoveries. We have to be willing to question our most deeply held assumptions, to see outside of them.

Really doing it

How can we really stay with these practices as we do this work? It is very helpful to have the support of a community that is aligned with these principles. There can be times when breaking out of our frames can lead to anxiety or other instability. The support of friends and community is essential in those moments to help reconnect us with what is real. At the same time, the right community can help us stay on target and with what really matters, whereas well-intentioned but misaligned friends can throw up new obstacles even as they try to help, as one has to shoulder the burden of staying on target themselves.

Community can also be very helpful in maintaining the form of the practices. It's hard to meditate regularly and harder to meditate to support work of this kind. Community practice and norms provide momentum that carries all. Supportive community questions us, helps us question ourselves, and provides a reflective mirror so we can see ourselves and our frames better. Community also provides useful guardrails that help us from going off the deep end in directions that aren't relevant, useful or beneficial, by providing real world interactions to ground us in the effects of the changes we go through.

Conclusion

We have described the value measurement problem and agency identification problem – both well-known within the AI alignment discourse – as "hard problems". We have argued that these problems are "hard" in a way that is quite distinct from other technical design problems. In the case of the value measurement problem, the difference is that other engineering problems pre-suppose some objective measure somewhere as a reliable measure of success, whereas in the value measurement problem we lack any such pre-supposable measure. In the case of the agency identification problem, the difference is that other characterization problems permit some kind of black-box assessment at some level of granularity, whereas the agency identification problem rules this out due to the possibility of deception. We have described hard problems in general, as well as their relationship to paradigm shifts, paradoxes, and koans. We have described pitfalls that arise

in practical work on hard problems, and the ways in which we believe that work can move forward in this area.