
GOAL

Deconfusing Goal-Directedness

Why You Should Care About Goal-Directedness

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Introduction

Deconfusing goal-directedness would boost your favorite research approach for solving AI Alignment.

Why? Because every approach I know of stands to gain from the clarification of goal-directedness, from Prosaic AGI Alignment to Agents Foundations. In turn, this ubiquitous usefulness of goal-directedness motivates the writing of this sequence, which will include a literature review of the idea in the AI Safety literature and beyond, as well as advanced explorations of goal-directedness by me and collaborators Michele Campolo and Joe Collman.

But before that, I need to back up my provocative thesis. This is why this post exists: it compiles reasons to care about goal-directedness, from the perspective of every research approach and direction I could think of. Although not all reasons given are equally straightforward, none feels outrageously far-fetched to me.

I thus hope that by the end of this post, you will agree that improving our understanding of goal-directedness is relevant for you too.

Thanks to Michele Campolo and Joe Collman for many research discussions, and feedback on this post. Thanks to Alexis Carlier, Evan Hubinger, and Jérémy Perret for feedback on this post.

Meaning of Deconfusion

Before giving you the reasons for caring about goal-directedness, I need to synchronize our interpretations of “deconfusion”. The term comes from MIRI, and specifically this blog post; it captures the process of making a concept clear and explicit enough to have meaningful discussions about it. So it’s not about solving all problems related to the concept, or even formalizing it perfectly (although that would be nice) -- just about allowing coherent thinking. To quote Nate Soares (MIRI’s Executive Director, and the author of the linked blog post):

> By deconfusion, I mean something like “making it so that you can think about a given topic without continuously accidentally spouting nonsense.”

What would that look like for goal-directedness? At first approximation, the idea simply means the property of trying to accomplish a goal. Which feels rather simple. But after digging deeper, issues and subtleties emerge: the difference between having a goal and being competent at accomplishing it (discussed here), what should count as a goal (discussed here), which meaningful classes of goals exist (discussed here), and many others.

Thus the concept is in dire need of deconfusion. Such clarification could take many forms, including:

- A mathematical formalization
- A decomposition into formalized components
- A decomposition into simpler and less confused informal components
- A list of accepted examples with different levels of goal-directedness
- A list of properties and their link with the intuitions behind goal-directedness
- And many more variants

Obviously, only time will reveal the form of our results on goal-directedness. Still, it's valuable to keep in mind the multitudes of shapes they could take.

Reasons To Care

Let's be honest: just listing research approach after research approach, and my reason for the relevance of goal-directedness to them, might be too much information to take in one reading. Fortunately, the reasons I found show some trends, and fit neatly into three groups.

- **(Overseeing)** In some cases, alignment comes from supervisors and overseers that monitor the AI during training. Goal-directedness is a natural and fundamental property to check, because of its many negative consequences. So deconfusion would facilitate checks of this important property by overseers and supervisors, and thus improve every approach depending on monitoring.
- **(Additional Structure on Utility/Reward Functions)** Many approaches to alignment rely on utility functions and reward functions to capture goals and values. Such representations are powerful, but so general that maximizing a utility function or a reward function doesn't reveal much about whether the system actually follows a goal or not (see the discussion here and here). Furthering our understanding of goal-directedness could reveal more structure to add on these representations of goals, making the pursuit of such a "goal" more closely tied to being goal-directed.
- **(Natural Mathematical Abstraction)** When attempting to formalize and clarify many aspects of decision making, AI and alignment, concepts like agency and optimization play a big role. Goal-directedness naturally relates to both, because agents are generally considered goal-directed, and so are explicit optimizers doing internal search. Thus goal-directedness should intuitively play a role in these formalizations, whether as a building block, a metric or an example to draw from.

Overseeing

The reasons from this section assume the use of an overseer. This is common for approaches affiliated with Prosaic AI Alignment, where the gist of alignment emerges from training constraints that forbid, push and monitor specific behaviors.

Interpretability and Formal Methods

Interpretability is one way to monitor an AI: it studies how the learned models work, and how to interpret and explain them. Similarly, formal methods (applied to AI) take

a formal specification, a model of computation and an AI, and verify whether the AI follows the specification when executed on this model of computation.

Ultimately, both interpretability and formal methods try to check properties of trained models, notably neural networks. Goal-directedness is an example of an important property to look for, as discussed above. And deconfusing goal-directedness would move us towards finding a specification of this property.

(Interpretability à la Clarity Team at OpenAI (for example here) might also prove important in deconfusing goal-directedness, by letting us look into and compare systems with various levels of goal-directedness)

IDA and Debate

Iterated Distillation and Amplification (IDA) and AI Safety via Debate (Debate) are two alignment schemes proposed respectively by Paul Christiano and Geoffrey Irving, and extended by many others.

IDA attempts to align a superintelligent AI by starting from a simple AI, amplifying it (training a second AI to imitate the human supervisor using the simple AI), and then distilling this amplified version (by training a simpler model to imitate the amplified AI) into a new AI that can be used by the human supervisor. Hopefully, repeating this will eventually create an AI with superhuman capabilities, while maintaining alignment.

Debate on the other hand places the human supervisor as a judge of a debate between two AIs. This value of debate comes from extending the reach of human feedback: judging a debate (which only presents bits and pieces of the arguments) is intuitively easier than checking a complete solution, which is easier than finding a solution. Whether or not debate works hinges on the importance of honesty for optimal play, and other theoretical and empirical questions about human evaluation of debates.

The current approach to IDA, and some approaches to Debate, rely on the same building block, factored cognition: answering questions by splitting them into subquestions given to other AIs with some but not all the context, with recursive splitting allowed. And a big question with this factorization is how to do it without allowing one subAI to manipulate the whole decision.

Deconfusing goal-directedness would help, in that checking the level of goal-directedness of the subAIs, or what kind of goals they might have, correlates with checking non-manipulation. Indeed, manipulation probably requires incentives to make long term plans towards something, which looks, smells and sounds like goal-directedness.

Ascription Universality

Universality (see also my own explanation) is a property of an overseer that matters a great deal to approaches like IDA and Debate. Basically, an overseer O is ascription universal for a computation C with regard to an epistemic viewpoint (usually the human supervisor H) if, whatever finite information about O and C is provided to H, the latter will privilege the probabilistic predictions of O over the ones of C. This is supposed to capture the fact that O is able to reveal the different beliefs or reasons that might be ascribed to C to explain it.

One very natural way to ascribe belief is through assuming goals and a desire to accomplish them. This is basically the intentional stance of Daniel Dennett, a classical framework to understand goal-directedness, and deconfusing goal-directedness would put such an ascription of beliefs into more formal and computational grounds.

Mesa-Optimization

The inner alignment problem, as stated in Risks from Learned Optimization, concerns learned optimizers: optimizers that result from an optimization process, like gradient descent. Such learned optimizers are called mesa-optimizers, and the inner alignment research direction studies two questions: under which conditions mesa-optimizers appear, and how aligned the mesa-objective is with the base objective of the initial optimization process.

One common criticism with mesa-optimizers is that their objective (the mesa-objective) is assumed to be explicitly represented in the system (see Richard Ngo's post here, for example). This is a bit weird, and looks more like a simplifying assumption than a property expected in actual systems. Moreover, maybe some of the reasoning about mesa-optimizers still works when considering learned goal-directed systems instead, as proposed in this comment by Tom Everitt.

By deconfusing goal-directedness, we might find properties of goal-directed systems and use them for studying the previous question: does the reasoning from Risks from Learned Optimization go through when considering learned models with these properties, instead of learned explicit optimizers? Such an investigation could shift the focus of research on inner alignment, or reinforce the importance of internal optimization.

Deceptive Alignment

One especially worrying case of mesa-optimization is deceptive alignment: a mesa-optimizer with a different objective from the base objective, but which is competent enough to deceive us into believing that it is pursuing the base objective. This can take many forms, from behaving nicely during training and defecting when put in the world, up to more outlandish ideas like gradient hacking.

The foremost proposals for dealing with deceptive alignment rely on relaxed adversarial training: using the overseer feedback (on the risk of catastrophic behavior) as part of the training signal. But getting this overseer feedback right, notably the detection of deceptiveness, proves difficult.

One possibility is to use myopia. Intuitively, myopia is supposed to capture the property that a system only makes short-term plans. Then the hope is that deceptive systems should probably be non-myopic. Thus we could detect non-myopia instead of deception, which is hopefully easier, and get the overseer feedback necessary for relaxed adversarial training.

Now, defining myopia is an open research problem -- see these two posts for pointers. Yet there is an interesting connection with goal-directedness: multiple researchers consider long-term goals as an important component of goal-directedness. Hence disentangling the various intuitions about goal-directedness could help deconfuse the idea of long-term goals, which in turn would help tremendously for deconfusing myopia.

Additional Structure on Utility/Reward Functions

Reasons in this section apply to a broader range of alignment proposals. Their common thread is to assume that utility functions or reward functions are used to capture goals and values.

Agent Incentives

The Safety Team at DeepMind wrote many different papers on agent incentives; specifically, on observation and intervention incentives that come from having a specific goal. Assuming a causal graph of the system and a goal, graphical criteria exists to find which nodes would be useful to monitor (observation incentives), and which nodes would be useful to control (intervention incentives). For goals, these papers consider controlling a utility node in the causal graph. That is, this research places itself within the framework of expected utility maximization.

As mentioned before, utility functions look too general to capture exactly what we mean by goals: every system can be seen as maximizing some utility function, even those intuitively not goal-directed. Deconfusing goal-directedness might allow the derivation of more structure for goals, which could be applied to these utility functions. The goals studied in this approach would then model more closely those of actual goal-directed systems, allowing in turn the derivation of incentives for more concrete and practical settings.

Value Learning

Value Learning is a pretty broad idea, which boils down to learning what we don't want the AI to mess up (our values), instead of trying to formalize them ourselves. This includes the reward modeling agenda at DeepMind, work on Cooperative Inverse Reinforcement Learning and Inverse Reward Design at CHAI, Stuart Armstrong's research agenda and G Gordon Worley III's research agenda, among others.

For all of these, the main value of deconfusing goal-directedness is the same: learning values usually takes the form of learning a utility function or a reward function, that is something similar to a goal. But values probably share many of the structure of goals. Such structure could be added to utility functions or reward functions to model values, if we had a better understanding of goal-directedness.

Impact Measures

Impact measures provide metrics for the impact of specific actions, notably catastrophic impact. Such an impact measure can be used to ensure that even a possibly misaligned AI will not completely destroy all value (for us) on Earth and the universe. There are many different impact measures, but I'll focus on Alex Turner's Attainable Utility Preservation (AUP), which is the one I know best and the one which has been discussed the most in recent years.

Attainable Utility Preservation ensures that the attainable utilities (how much value can be reached) for a wide range of goals (reward/utility functions) stays the same or improves after each action of the AI. This should notably remove the incentives for power-seeking, and thus many of the catastrophic unaligned behaviors of AI (while not solving alignment itself).

You guessed it, here too the value of goal-directedness comes from defining goals with more structure than simple utility or reward functions. Among other things, this might help extend AUP to more realistic environments.

Natural Mathematical Abstraction

Lastly, these reasons concern the Agents Foundations part of AI alignment research. They thus assume a focus on formalization, with applications to practical problems of alignment.

Mathematical Theory of RL and Alignment

Vanessa Kosoy from MIRI has been the main proponent of the creation of a mathematical theory of RL and alignment. Her point of view focuses on deriving formal guarantees about alignment in a learning theoretic setting, and this requires a theory of RL dealing with issues like non-realizability and traps.

Such guarantees will probably depend on the goal-directedness of the system, as different levels of goal-directedness should produce different behaviors. So knowing how to capture these levels will ground the dependency of the guarantees on it.

(Note that Vanessa already has her own definition of goal-directed intelligence, which doesn't seem to completely deconfuse goal-directedness, but may be sufficient for her research).

Embedded Agency

Embedded Agency is a broad class of research directions that focus on dealing with theoretical issues linked to embeddedness -- the fact that the AI inhabits the world on which it acts, as opposed to dualistic models in which the AI and the environment are cleanly separated. The original research agenda carves out four subproblems: Decision Theory, Embedded World Models, Robust Delegation and Subsystem Alignment. I'll focus on Embedded World Models, which has the clearest ties to goal-directedness. That being said, the others might have some links -- for example Subsystem Alignment is very close to Inner Alignment and Deceptive Alignment, which I already mentioned.

Embedded World Models ask specifically how to represent the world as a model inside the agent. Trouble comes from self-reference: since the agent is part of the world, so is its model, and thus a perfect model would need to represent itself, and this representation would need to represent itself, ad infinitum. So the model cannot be exact. Another issue comes from the lack of hardcoded agent/environment boundary: the model need to add it in some way.

Understanding goal-directedness would hopefully provide a representation of systems with goals in a compressed way. This helps both with the necessary imprecision of the map (notably because the AI can model itself this way) and to draw a line between such systems and the complex world they inhabit.

Abstraction

John S. Wentworth's research on abstraction centers around one aspect of Embedded World Models: what can be thrown out of the perfect model to get a simpler non-self-referential model (an abstraction) that is useful for a specific purpose?

Using goal-directedness for modelling systems in a compressed way is an example of a natural abstraction. Searching for a definition of goal-directedness is thus directly relevant to abstraction research, both because of its potential usefulness for building abstractions, and because it's such a fundamental abstraction that it might teach us some lessons on how to define, study and use abstractions in general.

Conclusion

To summarize, for a broad range of research agendas and approaches, deconfusing goal-directedness is at least partially relevant, and sometimes really important. The reasons behind that statement fit into three categories:

- Helping an overseer to check for issues during training
- Adding structure to utility functions/reward functions to make them behave more like goals.
- Abstracting many important systems into a compressed form..

So you should probably care about goal-directedness; even without working on it, taking stock of what has been done in this question might impact your research.

The next post in this sequence lay the groundwork for such considerations, by reviewing the literature on goal-directedness: the intuitions behind it, the proposed definitions, and the debates over the shape of a good solution to the problem.

Literature Review on Goal-Directedness

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Introduction: Questioning Goals

Goals play a central role in almost all thinking in the AI existential risk research. Common scenarios assume misaligned goals, be it from a single AGI (paperclip maximizer) or multiple advanced AI optimizing things we don't want (Paul Christiano's What Failure Looks Like). Approaches around this issue ask for learning the right goals (value/preference learning), allowing the correction of a goal on the fly (corrigibility), or even removing incentives for forming goals (CAIS).

But what are goals, and what does it mean to pursue one?

As far as we know, Rohin Shah's series of four posts were the first public and widely-read work questioning goals and their inevitability in AI Alignment. These posts investigate the hypothesis that goals are necessary, and outline possible alternatives. Shah calls the property of following a goal "goal-directedness"; but he doesn't define it:

> I think of this as a concern about *long-term goal-directed behavior*. Unfortunately, it's not clear how to categorize behavior as goal-directed vs. not. Intuitively, any agent that searches over actions and chooses the one that best achieves some measure of "goodness" is goal-directed (though there are exceptions, such as the agent that selects actions that begin with the letter "A"). (ETA: I also think that agents that show goal-directed behavior because they are looking at some other agent are not goal-directed themselves -- see this comment.) However, this is not a necessary condition: many humans are goal-directed, but there is no goal baked into the brain that they are using to choose actions.

Later on, he explains that his "definition" of goal-directedness relies more on intuitions:

> Not all behavior can be thought of as goal-directed (primarily because I allowed the category to be defined by fuzzy intuitions rather than something more formal)

Clearly, fuzzy intuitions are not enough to decide whether or not to focus on less goal-directed alternatives, if only because we can't define the latter. We thus need a more advanced understanding of the concept. What's more, deconfusing this concept would probably move forward many existing research approaches, as defended in the first post of this sequence.

That being said, intuitions shouldn't be thrown out of the window either. As suggested by Vanessa Kosoy, they provide the analogue of experimental data for philosophy: theories and definitions should mostly agree with intuitions in the simple and obvious cases. That doesn't mean we should be slaves to intuitions; just that biting the bullet on breaking one of them asks for a solid foundation about most other basic intuitions.

Hence this literature review. In it, we go through the most salient and important intuitions about goal-directedness we found in the literature. We do not limit ourselves to AI Alignment research, although many references come from this field. The end goal is to crystallize tests we can use on proposals for goal-directedness, a clear benchmark against which to judge proposals for deconfusing goal-directedness.

Note that we don't necessarily share the intuitions themselves; instead, we extract them from the literature, and delay further exploration to subsequent posts in this sequence.

This post follows a three part structure:

1. **(Intuitions about Goal-Directedness)** The first and most relevant part explores five topics related to goal-directedness we found again and again in our readings, and extracts a test for each to "falsify" a proposal for goal-directedness.
2. **(Comparison with Optimization and Agency)** The second asks about the difference between goal-directedness and the intertwined concepts of optimization and agency.
3. **(Proposals for Goal-Directedness)** The third and last part studies some proposals for a definition of goal-directedness, seeing how they fare against the tests extracted from the first part.

Note that, as mentioned in our previous post in this sequence, clarification of goal-directedness could take many forms. It seems improbable to be usefully characterised as a binary property, but it's not clear whether it should be seen as a continuum, a partial order, or something else entirely. This post doesn't assume anything except that there are different levels of goal-directedness, with some comparable and some maybe not.

Thanks to Evan Hubinger, Richard Ngo, Rohin Shah, Robert Miles, Daniel Kokotajlo, and Vanessa Kosoy for useful feedback on drafts of this post.

Intuitions about Goal-Directedness

In this section, we explore five main topics related to goal-directedness from the literature. These are:

- **(What Goals Are)** We discuss the possible definitions of goals, focusing on utility functions and their variants. We also explore some proposed alternatives.
- **(Explainability)** We unravel the link between goal-directedness and explainability of the system in terms of goals. This includes both which forms of explainability are considered, and their effectiveness.
- **(Generalization)** We study the link between goal-directedness and generalization, notably the apparent consensus that goal-directedness directly implies some level of generalization.
- **(Far-sighted)** We study the link between goal-directedness and the timescale over which the impacts of actions are considered. Although not common in the more general literature, this connection appears in almost every resource considered in the AI Alignment literature.
- **(Competence)** We study the link between goal-directedness and the ability of the system to accomplish a given goal and/or to be efficient in accomplishing

this goal. The distinction appears relevant as most resources don't posit the same link between these two forms of competence and goal-directedness.

What Goals Are

Goals are what goal-directed systems push towards. Yet that definition is circular because we don't know how to formalize goal-directed systems! One way to break the circle is to explore goals themselves; we believe this is paramount for deconfusing the idea of goal-directedness.

So let's start our investigation of the literature by looking at the answers to this question: what are goals? Or more precisely, what is the space of all possible goals?

Thinkers with a decision theory mindset might answer utility functions: functions from states of the world, or histories of such states, to real numbers, capturing preferences. Indeed, a classic argument in the AI Alignment literature defends that a future superintelligent AI could be modeled as maximizing expected utility, because otherwise it would reveal incoherent preferences, and thus it could be exploited by humans, in contradiction with the assumption of superintelligence. Given the value for AI Alignment in modeling superintelligent AI, this would be strong evidence for the… utility of utility functions as a formalization of goals.

Yet the story doesn't stop there: Shah's most discussed post on goal-directedness, Coherence arguments do not imply goal-directed behavior, argues that the class of utility functions is too large: in principle, any behavior can be interpreted as maximizing some utility function, even ones that are intuitively not goal-directed. This comes from considering a utility function that returns 1 for the exact histories or states appearing in the behavior, and 0 for the rest.

Did Shah's arguments convince AI Alignment researchers? Looking at the top comments on the post, it seems so. Moreover, the only pushback we were able to find on this post concerns the interpretation of the coherence arguments or Shah's conclusions (like Wei Dai's post or Vanessa Kosoy's review), not the criticism of utility functions as representing goals.

There is a fleshing out of the reasoning in Richard Ngo's Coherent behaviour in the real world is an incoherent concept. Ngo does so by exploring two definitions of coherence (one with preference over states, the other with preferences over state trajectories), and considering how to use the formal coherence results on such definitions for more practical settings. The conclusion? Whichever way we use them, mathematical coherence arguments can only be applied non-trivially to real-world context by further constraining preferences to what we humans consider relevant. At which point Ngo argues that talking of goal-directedness is just better, because it acknowledges the subjective human perspective element instead of sweeping it under the rug of mathematical elegance.

If not utility functions, then what? Shah mentions that a simple utility function (the kind that can be explicitly represented in a program) would probably lead to goal-directedness.[^1]

> As a corollary, since all behavior can be modeled as maximizing expected utility, but not all behavior is goal-directed, it is not possible to conclude that an agent is goal-driven if you only know that it can be modeled as maximizing some expected

> utility. However, if you know that an agent is maximizing the expectation of an *explicitly represented* utility function, I would expect that to lead to goal-driven behavior most of the time, since the utility function must be relatively simple if it is explicitly represented, and *simple* utility functions seem particularly likely to lead to goal-directed behavior.

Comment on Coherence arguments do not imply goal directed behavior by Ronny picks this line of thought, focusing implicitly on simpler utility functions as goals. This post defends that the real question is whether the expected utility maximizer is the best/simplest/most efficient explanation. Indeed, many of the cases where the utility maximization is trivial come from the utility function capturing all the complexity of the behavior forever, which is probably just as complex an explanation as the mechanical one. So the utility functions to consider are probably simple/compressed in some way.

The emphasis on simplicity also fits with a big intuition about when goal-directedness might appear. For example, Risks from Learned Optimization, by Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant, argues that the push towards simple models incentivizes the selection of optimizers as learned models, where optimizers are the representative of high goal-directedness in this paper.

Beyond simplicity, Shah's quote above pushes to further constrain the utility functions considered. Hoagy's When do utility functions constrain? explicitly studies this option. The post proposes a constraint on the class of utility considered:

> What we need to find, for a given agent to be constrained by being a 'utility maximiser' is to consider it as *having a member of a class* of utility functions where the actions that are available to it systematically alter the expected utility available to it - for **all** utility functions within this class.

It tries to capture the fact that intuitively goal-directed systems will attempt to shift the expected utility where they want, instead of having the maximal expected utility given to them from the start.

Another possibility[^2] would be to ditch utility functions completely. What might this look like? The most explicit take on that comes from Ngo's AGI safety from first principles: Goals and Agency. Among other things, this post argues for goals as functions of the internal concepts of the system, instead of functions of behaviors or states of the world.

> Given the problems with the expected utility maximisation framework I identified earlier, it doesn't seem useful to think of goals as utility functions over states of the world. Rather, an agent's goals can be formulated in terms of whatever concepts it possesses - regardless of whether those concepts refer to its own thought processes, deontological rules, or outcomes in the external world.

The issue of course is that finding such concepts (and even defining what "concept" means) is far from a solved problem. Interpretability work on extracting meaning from trained models might help, like the circuit work of the Clarity team at OpenAI.

So concept-based definitions of goals are hard to investigate. But some parallel experimental work on goal-inference gives evidence that goals based more on logical concepts and relations can be used concretely. Tan Zhi-Xuan, Jordyn L. Mann, Tom Silver, Joshua B. Tenenbaum, and Vikash K. Mansinghka's Online Bayesian Goal Inference for Boundedly-Rational Planning Agents studies the goal-inference problem

for systems, expressing goals in PDDL, a specification language for planning tasks and environments. Their inference algorithm performs better than standard Bayesian IRL, and shows impressive similarities with human inferences for the same trajectories.

Lastly, we would like to touch on one last interesting idea that appears important to the question of defining goals. It comes from this quote from Shah's Intuitions about goal-directed behavior.

> There's a lot of complexity in the space of goals we consider: something like "human well-being" should count, but "the particular policy <x>" and "pick actions that start with the letter A" should not. When I use the word goal I mean to include only the first kind, even though I currently don't know theoretically how to distinguish between the various cases.

Such a distinction intuitively separates goals about the state of the world, and goals about the output of the agent. But there seems to be more subtlety involved than just these two categories (as Shah himself acknowledges in discussing another example of non-goal for him, twitching).[^3]

This part of the literature thus suggests to us the following test for proposals of goal-directedness: **goals cannot just be the set of all utility functions. They must either be more constrained sets of utility functions, or different altogether, like concept-based goals.**

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Explainability

Why do we think so much about goals when looking at the world and trying to understand it? Why do we value thinking in terms of goals, and apply such thinking to almost anything that crosses our mind?

One answer is because it works pretty well. If a system is well described as pursuing a goal, then it will probably do things that we can interpret as bringing it closer to its goal. To take a classic example, modeling AlphaGo as wanting to reach the goal of winning does help in predicting, if not its exact next move, at least that the outcome of the game, or the fact it won't play terrible moves. Even when predicting what a human would do against AlphaGo (a situation where AlphaGo will almost certainly win), knowing that the human has the goal of winning helps us interpret why they're playing these specific moves.

The thing is, this notion of explainability becomes more subtle if the system is not as competent as AlphaGo. But we defer a discussion of competence in the context of goal-directedness to the relevant subsection below. Until then, let's assume that the system is competent enough to be explained as having a goal, if it has one.

Such a setting is exactly the one of The Intentional Stance by Daniel Dennett, often cited when discussing goals in AI Alignment research. In order to capture the different approaches one can take to model a system, Dennett defines three stances:

- The physical stance, which models the system as the sum of its physical components, and predicts its behavior by simulating the laws of physics.

- The design stance, which models the system as resulting from a design (either by an intelligent designer or by a process like evolution), and thus as made of higher level parts with each a reason for existing and interacting as they do. For example, considering an electronic circuit as made of logic gates is using the design stance, because it abstracts away the physical details that are not relevant to the design at hand. (This corresponds to having a gears-level model)
- The intentional stance, which models the system as having beliefs (models of the world) and desires (goals), and as rationally trying to reach its desires by acting in accordance to its beliefs. The classical example is humans (Dennett's first goal with the intentional stance is to explain folk psychology, the intuitive understanding of how other people think and react), but Dennett applies this stance to many other systems, including thermostats!

Given a system, which stance should we apply to it? Dennett doesn't say explicitly, at least in The Intentional Stance. But he suggests that going from physical to design, or from design to intentional, is only justified if it improves our predictive powers (in terms of accuracy and/or efficiency). Applied to humans, Dennett argues that most of the time this means using the intentional stance, because it is the only tractable stance and it usually works.

He even goes one step further: in the context of theory of mind, he defends that the true believers (systems that should be ascribed internal states like beliefs) are exactly the intentional systems. Given how the intentional stance fits with the idea of goal-directedness, we can extend his ideas to say that the goal-directed systems (those that should be ascribed a goal) are exactly the intentional systems. What happens inside doesn't matter as long as the intentional stance works well enough.

In a similar vein, in his book Life 3.0, Max Tegmark gives a definition of goal-directedness (called goal-oriented behavior) that depends heavily on explainability:

> Goal-oriented behavior: Behavior more easily explained via its effects than via its cause

This approach fits with Dennett's intentional stance when "cause" is understood as the mechanistic cause, while "effects" is understood as the end result. Max Tegmark never explicitly defines these terms, but his writing on goal-oriented behavior implies the ones we gave.

> Consider, for example, the process of a soccer ball being kicked for the game-winning shot. The behavior of the ball itself does not appear goal-oriented, and is most economically explained in terms of Newton's laws of motion, as a reaction to the kick. The behavior of the player, on the other hand, is most economically explained not mechanistically in terms of atoms pushing each other around, but in terms of her having the goal of maximizing her team's score.

Such approaches deal with the issue raised by Shah about the vacuousness of utility maximization. Recall that every system can be thought of as maximizing expected utility for some well-chosen utility functions. But Dennett (and Tegmark) doesn't only require that systems are explained by the intentional stance (as everything can be); he also wants the explanation to be better in some sense than mechanical ones. So the vacuousness of the expected utility maximization is not important, as we only want to consider the systems best explained in this way. It's also the interpretation favored by Ronny, on this answer to Shah.

There is also an argument to be made that goal-directedness is directly related to the kind of explanations that humans find intuitive. As we wrote above, Dennett's intentional stance attempts to explain folk psychology, which is pretty much this ability of humans to derive useful explanations *of other human beings*. So creating goal-directed systems would leverage these systems for prediction, as explained in Shah's Will humans build goal-directed agents?:

Another argument for building a goal-directed agent is that it allows us to predict what it's going to do in novel circumstances. While you may not be able to predict the specific actions it will take, you can predict some features of the final world state, in the same way that if I were to play Magnus Carlsen at chess, I can't predict how he will play, but I can predict that he will win.

That being said, Shah thinks that prediction can be improved by going beyound goal-directedness:

> I also think that we would typically be able to predict significantly *more* about what any AI system we actually build will do (than if we modeled it as trying to achieve some goal). This is because "agent seeking a particular goal" is one of the simplest models we can build, and with any system we have more information on, we start refining the model to make it better.

Nonetheless, some experimental work on explaining systems with the intentional stance partly vindicates Dennett.

As already mentioned, Zhi-Xuan et al.'s Online Bayesian Goal Inference for Boundedly-Rational Planning Agents shows how useful explanations of behaviors can be extracted through goal-inference. This gives evidence for the predictive power of this stance. Another evidence is arguably the whole field of Inverse Reinforcement Learning.

In Agents and Devices: a Relative Definition of Agency, Laurent Orseau, Simon McGregor McGill and Shane Legg compare mechanistic (a mix of physical and design stance) explanations against intentional explanations of trajectories in toy environments. The devices (for mechanistic explanations) are functions taking a history and returning a probability distribution over actions; the agents (for intentional explanations) use Bayesian RL. Through an appropriate simplicity prior, and the use of Inverse Reinforcement Learning for the agent side, Orseau et al. can compute a posterior distribution over devices and agents as explanations of a given trajectory. Their results fit with intuition: running in circles is inferred to be device-like, because it has a simple mechanical explanation; going straight to a specific zone is inferred to be agent-like; and others in between.

Lastly, we want to address a comment by Shah on a draft of this post. He wrote

> I think this section is a reasonable way to talk about what humans mean by goals / goal-directed behavior, but I don't think the resulting definition can tell us that much about AI risk, which is what I want a definition of goal-directedness for.

Although this goes beyond the subject of this literature review, one way in which more explainable systems might be more dangerous in terms of AI risk is that they will behave in ways that follow fundamental intuition in humans. Hence when we find a scenario involving such systems, we should put more credence on it *because* it applies to systems for which our intuitions are better calibrated by default. This of

course doesn’t mean that less explainable systems are not also potential sources of AI risk.

To conclude, this part of the literature suggests to us the following test for proposals of goal-directedness: **a definition of goal-directedness should come with a way to explain systems based on goals (like saying they maximize some utility). And the predictive power (both in terms of efficiency and in terms of precision) of this explanation compared to purely mechanistic ones should increase with goal-directedness.**

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Generalization

Maybe the most obvious aspect of goal-directedness is that it implies generalization: a system striving towards a goal will adapt itself to changes in the environment.

Indeed, this distinction between goal-directed behavior and the opposed habitual behavior has been a major topic in psychology and cognitive neuroscience for more than 50 years. Goals and Habits in the Brain, by Ray J. Dolan and Peter Dayan, surveys this research, splitting the approaches to the goal-directed/habitual dichotomy into four overlapping “generations”: the behavioral approach in rodents; the behavioral approach extended to humans; a computational modeling approach through model-based vs model-free RL; and the refinement of this modeling. As they write, habitual behavior is understood as automatic, and thus efficient but inflexible, whereas goal-directed behavior is understood as thoughtful, and thus expensive but flexible.

> Thus, key characteristics of habitual instrumental control include automaticity, computational efficiency, and inflexibility, while characteristics of goal-directed control include active deliberation, high computational cost, and an adaptive flexibility to changing environmental contingencies.

Note that the current consensus presented in this survey argues that both behaviors play a role in human cognition, with a complex dynamic between the two. Yet time and time again, in experiment after experiment, goal-directed behavior becomes necessary when something changes that invalidates the previous habits.

Similarly to this modeling of behavior, Dennett’s intentional stance assumes that the system will alter its behavior in response to changes in the environment. Dennett even uses the sensitivity of this change as a means for comparing two intentional systems.

> But as systems become perceptually richer and behaviorally more versatile, it becomes harder and harder to make substitutions in the actual links of the system to the world without changing the organization of the system itself. If you change its environment, it will notice, in effect, and make a change in its internal state in response. There comes to be a two-way constraint of growing specificity between the device and the environment. Fix the device in any one state and it demands a very specific environment in which to operate properly (you can no longer switch it easily from regulating temperature to regulating speed or anything else); but at the same time, if you do not fix the state it is in, but just plonk it down in a changed environment, its sensory attachments will be sensitive and discriminative

> enough to respond appropriately to the change, driving the system into a new state, in which it will operate effectively in the new environment.

Generalization also plays a fundamental role in the analysis of goal-directedness in the AI Alignment literature. In Intuitions about goal-directed behavior, Shah even defines goal-directedness that way:

> This suggests a way to characterize these sorts of goal-directed agents: there is some goal such that the agent's behavior *in new circumstances* can be predicted by figuring out which behavior best achieves the goal.

We failed to find any rebuttal to the claim that goal-directedness would entail generalization, at least with some assumption of competence.[^4] On the other hand, the arguments in favor abound:

- Shah's Will humans build goal-directed agents? presents the argument that goal-directed systems are the best known candidate for solving problems in novel and possibly superhuman ways, which amount to generalizing beyond our current context.
- Ngo's AGI safety from first principles: Goals and Agency lists criteria for goal-directedness, and the last one, flexibility, directly references the ability to adapt plans to changes in circumstances.
- Hubinger et al.'s Risks from Learned Optimization uses explicit optimisers as a representation of goal-directed systems, both because of the usefulness of knowing the internal structure, and because such optimizers will generalize better. That being said, the risk of inner alignment is that such a learned optimizer becomes goal-directed towards a different goal than the expected one from training. So in some sense, while goal-directed systems are assumed to generalize, mesa-optimizers present the risk of undesirable generalization.
- Wei Dai's Three ways that "Sufficiently optimized agents appear coherent" can be false lists distributional shift as one way that an optimized system might appear incoherent, even if it's performing well on the training environments. In contraposition, a truly coherent system (a system which always appears to follow some goal) would deal with this distributional shift, and thus generalize.

This part of the literature thus suggests to us the following test for proposals of goal-directedness: **generalization should increase with goal-directedness, for a suitable definition of generalization that captures adaptation of behavior to changes in the environment.**

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Far-sighted

One intuition about goal-directedness mostly appears in the AI Alignment literature: that goal-directed systems should consider the long-term consequences of their actions. Not that such far-sightedness is forbidden in goal-directed behavior from cognitive neuroscience, or in Dennett’s intentional stance. But it holds no special place: a system that only looks at short-term consequences of its actions would still be considered goal-directed or intentional. In AI Alignment on the other hand, researchers who looked into the issue apparently agree that far-sightedness plays a fundamental role in goal-directedness.

Look for example at Ngo's AGI safety from first principles: Goals and Agency: one of his criteria for goal-directedness is large scale. Or see Shah's list of non-goal-directed AI designs: it includes act-based agents whose main feature is to only care about the short term preferences of the human supervisor/AI. Or observe that Eric Drexler's CAIS are presented as different from goal-directed systems in part because they only have the resources to consider short-term consequences of their actions.

Why is it so? It probably stems from safety issues like Stephen Omohundro's Convergent Instrumental Subgoals, Hubinger et al.'s Deceptive Alignment, and Goodhart's Law. The first two require the consideration of non-immediate outcomes, and the third only really becomes an existential problem when considering large timescales where enormous optimization pressures are applied.

So this intuition is clearly relevant for thinking about AI Alignment and existential risks. Let's just keep in mind that it doesn't necessarily apply to the general idea of goal-directedness as considered outside AI Alignment.

This part of the literature thus suggests to us the following test for proposals of goal-directedness: **the timescale of goals considered should increase with goal-directedness.**

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Link with Competence

Finally, the most subtle intuition about goal-directedness concerns its link with competence. Note that we're not discussing which goals a system will have (the sort of question addressed by Nick Bostrom's Orthogonality Thesis); instead, we focus on how competence changes with goal-directedness. This also means that we don't delve into whether competence pushes towards goal-directedness. The latter is an important question, studied for example in Shah's Coherence arguments do not imply goal-directed behavior and in David Krueger's Let's Talk about "Convergent Rationality"; but it is not our focus here.

So what competence is needed for goal-directedness? Let's first clarify what we mean by competence. We separate how far the system can go in accomplishing its goal (**ideal accomplishment**)[^5] from how fast it moves towards the goal (**efficiency**). When thinking about reaching a certain state, like winning at chess, ideal accomplishment measures the ability to win, possibly against a range of different adversaries, whereas efficiency measures things like the number of moves taken or how close were the won games. When thinking more about maximizing some utility in an online setting, ideal accomplishment talks about how little regret is possible for the system, whereas efficiency measures the actual regret of the system.

This split also manifests itself in the link with goal-directedness: ideal accomplishment is mostly considered independent of goal-directedness[^6], while efficiency is assumed to grow with goal-directedness.

Beyond a minimal level of ideal accomplishment (if the system is too incompetent, its goal cannot be revealed), behavioral approaches to goal-directedness like Dennett's intentional stance don't assume the ability to reach goals. The difference between a

good chess player and a bad chess player is not seen as a difference in their intentionality, but as a difference in their beliefs (models of the world).

As for structural and mechanical definitions, they don't even need this minimal assumption of ideal accomplishment: goal-directedness doesn't have to be revealed. Hence no criterion requires the system to be good at accomplishing its goal in Ngo's AGI safety from first principles: Goals and Agency, and it's not assumed either in Hubinger et al.'s Risks from Learned Optimization.

This assumption, that only minimal ideal accomplishment matters for goal-directedness, has some experimental evidence going for it. In Zhi-Xuan et al.'s Online Bayesian Goal Inference for Boundedly-Rational Planning Agents, the goal-inference assumes only bounded-rationality from the system (which can be seen as a constraint on ideal accomplishment). Even with this limitation, the authors manage to infer goals (among a small set of predefined ones) in some cases. More generally the Inverse Reinforcement Learning (IRL) literature contains many examples of goal-inference even with less than competent demonstration. For a recent example, see Jonnavittula and Losey's I Know What You Meant: Learning Human Objectives by (Under)estimating Their Choice Set: in it, they avoid overfitting the reward to the demonstrator's mistakes by making the inference more risk-averse, which means assuming that the demonstrator was limited to trajectories very close to the one taken.

What happens for our other facet of competence, efficiency? There, behavioral definitions do assume that efficiency grows with goal-directedness. The intentional stance for example assumes rationality; this means that a system well explained intentionally will move as efficiently as possible (given its beliefs) towards its goal. The same can be said for proposals related to Dennett's, like Ronny's rescue of utility maximization or Tegmark's definition of goal-oriented behavior.

While structural definitions don't explicitly require efficiency, their reliance on mechanisms like optimization, which approximate the ideal rationality used by Dennett, hint at the same dependence between efficiency and goal directedness.

That being said, one might argue that the link between efficiency and goal-directedness relates more to the alignment issues linked with goal-directedness than to the property itself. Jessica Taylor's quantilizers can be interpreted as such an argument. Quantilizers don't just maximize expected utility; they instead randomly sample from the top x% actions according to a utility function and a base distribution. The paper proposes that by doing this, the system might still accomplish the wanted goal without some of the expected consequences of stronger optimization.

> Given that utility maximization can have many unintended side effects, Armstrong, Sandberg, and Bostrom[12] and others have suggested designing systems that perform some sort of "limited optimization," that is, systems which achieve their goals in some non-extreme way, without significantly disrupting anything outside the system or otherwise disturbing the normal course of events. For example, intuition says it should be possible to direct a powerful AI system to "just make paper clips" in such a way that it runs a successful paper clip factory, without ever attempting to turn as much of the universe as possible into paper clips.

Rephrased in our words, quantilizers aim at reducing efficiency while maintaining ideal accomplishment. The assumption that this will keep accomplishing the goal in many cases thus undermines the dependence between efficiency and goal-directedness.

This part of the literature thus suggests to us the following test for proposals of goal-directedness: **ideal accomplishment (as in being able to accomplish the goal in many contexts) should only be relevant to goal-directedness in that a minimal amount of it is required. Efficiency on the other hand (as in measuring how fast the system accomplishes the goal) should increase with goal-directedness.**

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Comparison with Optimization and Agency

Why not use an already existing concept in place of goal-directedness? After all, we have two closely related alternatives at hand: optimization and agency. It might seem at first glance that goal-directedness merely measures optimization power or agency. The goal of this section is to clarify the differences, and justify an independent study of goal-directedness.

Optimization

What's the difference between goal-directedness and optimization? After all, many of the paper mentioned in the previous section either assume optimization (Hubinger et al. in Risks from Learned Optimization and Ngo in AGI safety from first principles: Goals and Agency) or use it as an implicit model of goal-directedness (Shah's Intuitions about goal-directed behavior).

Yet there is some sense in which goal-directedness appears different from optimization. To see that, we will first rephrase two definitions of optimization from the literature: internal search and optimizing systems.

- **(Internal Search)** This is the meaning from optimization algorithms: search a space of possibilities for the best option according to some objective (like training neural nets by gradient descent). It fits with Abram Demski's selection.
- **(Optimizing Systems)** This definition comes from Alex Flint's the ground of optimization: an optimization system robustly pushes the state of the system into a smaller set of target states. Note that this is different from Abram Demski's control: a control process only has one shot at doing what it wants to do, and doesn't have an explicit access to a search space (like a growing plant or a guided missile). Optimizing systems appear to encompass internal search/selection as well as control.

Armed with these two definitions, it's easier to compare with goal-directedness. Ideally, we would want to find a simple relation, like an equality or straightforward inclusions. Alas the literature is far from settled on such a proper hierarchy.

Let's start with internal search. Proponents of a structural definition of goal-directedness usually assume optimization in this sense, like Hubinger et al. in Risks from Learned Optimization and Ngo in AGI safety from first principles: Goals and Agency. So in a sense, they assume that goal-directedness is contained in internal search. One way to interpret the additional constraints they bring to optimization (for

example the self-awareness criterion explicitly stated by Ngo) is that goal-directed systems are a subset of systems using internal search.

On the other hand, proponents of behavioral definitions seem keen to accept non-optimizing systems. Recall that Dennett accepts a thermostat in the club of intentional systems -- and a thermostat definitely doesn't run any internal search.

Hence without rejecting one half of the literature, there is no way to include internal search in goal-directedness, or the other way around.

What about optimizing systems? At first glance, it seems that the situation is more straightforward: Alex Flint himself includes goal-directed systems inside the set of optimizing systems:

> Our perspective is that there is a specific class of intelligent systems — which we call optimizing systems — that are worthy of special attention and study due to their potential to reshape the world. The set of optimizing systems is smaller than the set of all AI services, but larger than the set of goal-directed agentic systems.

Yet the issue of competence muddles it all. Indeed, optimizing systems **robustly** push the configuration space towards a smaller target. That is, they are *successful* optimizers. To the extent that goal-directed systems don't have to be very competent (in reachability), they might fall short of forming optimizing systems with their environment.

And it is not the other way around either: some examples of optimizing systems (computing the square root of 2 through gradient descent or a tree growing) hardly satisfy all the tests for goal-directedness. So goal-directedness doesn't contain all optimizing systems.

All in all, the links between optimization (for either definition) and goal-directedness are far from obvious. Yet a general trend emerges from studying the literature: **most authors agree that goal-directedness isn't just optimization.**

Agency

Agency is a broad concept that appears in various fields, from psychology and economics to the natural sciences and AI. It is thus not surprising that different definitions of agency have already been proposed. In Defining Agency: individuality, normativity, asymmetry and spatio-temporality in action, Xabier Barandiaran, Ezequiel Di Paolo, and Marieke Rohde collect the insights found in the literature on agency and try to give a definition useful to multiple domains. According to them:

> [...] agency involves, at least, a system doing something by itself according to certain goals or norms within a specific environment.
>
> In more detail, agency requires:
>
> 1. Individuality: the system is capable of defining its own identity as an individual and thus distinguishing itself from its surroundings; in doing so, it defines an environment in which it carries out its actions;
> 2. Interactional asymmetry: the system actively and repeatedly modulates its coupling with the environment (not just passive symmetrical interaction);

3. Normativity: the above modulation can succeed or fail according to some norm.

Readers interested in more formal approaches to agency might consider Semantic information, autonomous agency and non-equilibrium statistical physics by Kolchinsky and Wolpert, or Martin Biehl's PhD thesis (and referenced papers).

Back to goal-directedness, the part of the definition by Baradiaran et al. that relates to our review is normativity as a necessary requirement: every agent acts according to some norm or goal, in a quite broad sense. They consider minimal proto-organisms as agents that modulate their interaction with the environment through processes aimed at dissipation and self-maintenance; but they classify a human undergoing involuntary tremors as non-agentic, since it is hard to find a sense in which tremors succeed, or fail, to fulfill any norm generated by the human system and related to its interaction with a generic environment.

The authors also write that

> [...] systems that only satisfy constraints or norms imposed from outside
>
> (e.g. optimization according to an externally fixed function) should not be treated as models of agency).

It is clear that Baradiaran et al. use the terms "agent" and "goal" differently than how they are usually used in AI: in the standard AI textbook Artificial Intelligence: A Modern Approach, Russell and Norvig present different types of agents, but classify only some of them as goal-based.

Dennett adopts another viewpoint: he argues for "the unavoidability of the intentional stance with regard to oneself and one's fellow intelligent beings" (The Intentional Stance, p.27); in other words, he claims that intelligent agents necessarily behave according to the beliefs-goals model.

On the other hand, when considering the arguments on AI risk, Ngo doesn't take for granted that AGI will be goal-directed. At the same time, in his analysis he uses the terms "agency" and "goal-directedness" interchangeably, interpreting them as two different terms for the same concept. He acknowledges that multiple factors could drive the development of goal-directed AGI, and he also points out that a collective intelligence as a whole could be more agentic than any individual agent within it.

Goal-directedness in distributed agent systems is also considered by Magnus Vinding in his short book Reflections on Intelligence. He notes that the collective human ability to achieve goals is great thanks to the wide variety of specialized tools that humans have created, and that its nature is highly distributed. Similarly, in CAIS Drexler argues that superintelligence won't necessarily be embodied in a single monolithic agent: he actually claims that distributed superintelligence will be obtained before single-agent AGI is developed.

Lastly, the philosophically-inclined reader can check the SEP page on Agency, which presents the standard view on the concept of action, relying on intentional actions and acting for a reason, and different kinds of agency. The page on Moral Motivation discusses instead Humeanism and Internalism/Externalism, which are different philosophical positions regarding the connection between beliefs and goals—both concepts are often used to describe the behavior of agents.

In summary: **In the literature we found significant, though far from perfect, overlap between agency and goal-directedness. Yet goal-directedness appears like the most tractable concept for the moment, compared to agency. High goal-directedness and agency can emerge also in distributed systems of agents that are, individually, less goal-directed.**

Proposals for Goal-Directedness

At last, let's see how the different proposals for goal-directedness fare for our criteria. We focus on the following four:

- Dennett's intentional stance[^7]
- The operationalization of the intentional stance with expected utility maximization
- Ngo's six criteria for goal-directedness
- Vanessa Kosoy's definition of goal-directed intelligence

Intentional Stance

Recall Dennett's intentional stance: an intentional system is the kind of system that can be more easily explained as following its desires (goals) according to its beliefs (models of the world). The comparison is drawn with the physical stance (describe the physical configuration of atoms, and simulate the laws of physics) and the design stance (assume that the system was designed and abstract it into a gears-level model).

How does it fare against our tests from the literature?

- **(What goals are)** Here the intentional stance disappoints: it says nothing about what is a goal, and what makes a goal. It relies on the common sense definition of a goal as a desire, which might suffice for folk psychology, but doesn't cut it for thinking about goals more broadly.
- **(Explainability)** The intentional stance is strong with this one -- after all explainability is in the definition! Arguments can be made about how much predictive power does the intentional stance give, but explaining the behavior of an intentional system is its point.
- **(Generalization)** The intentional stance also fares well on generalization, thanks to its focus on adaptability. Dennett is pretty clear that with a rich enough connection to the world, the beliefs should become accurate enough to adapt the behavior to new situations.
- **(Far-sighted)** As mentioned in the corresponding section, approaches to goal-directedness outside AI Alignment rarely consider far-sightedness as fundamental in the definition of goal-directedness. It thus makes sense that Dennett doesn't expand on this point.
- **(Link with Competence)** Ideal accomplishment seems irrelevant to the intentional stance except for the minimal level that allows the system to be efficiently predicted as an intentional system. Then efficiency and applicability of the intentional stance go hand in hand, as a more efficient system will be closer to Dennett's assumption of rationality.

All in all, the intentional stance scores decently enough on the tests we extracted from the literature (maybe in part because Dennett's book The Intentional Stance is such a big part of this literature). Yet this overlooks the main issue with the intentional stance: its lack of formalization. Yes, an intentional system is one that is better explained by the intentional stance; but what does "better explained" means here? Is it about computational complexity? About accuracy? About both? How many errors until the explanation is deemed not good enough? And how do we formalize this explanation, by the way? Utility maximizers are an obvious answer, but they have problems of their own, as we'll see for the next proposal.

A more formal grounding is important because of the ease with which we as humans use a version of the intentional stance to explain almost everything. A well-known example is ELIZA, the computer program that acted as a psychologist by recognizing specific keywords and rephrasing as questions what was written to it. This program had an incredibly basic design stance explanation, but users got emotionally attached to ELIZA really quickly (prompting the invention of the ELIZA effect). To quote ELIZA's programmer, Joseph Weizenbaum:

> What I had not realized is that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people.

On the other hand, a recent exploration of this issue, Do We Adopt the Intentional Stance Toward Humanoid Robots? by Marchesi et al., found out that participants tended to slightly favour the design stance when explaining the actions of a humanoid robot. Nonetheless, some situations can create mentalistic assumptions coherent with the intentional stance. This might mean that the ELIZA effect is mitigated when looking at actions specifically, instead of interacting directly with the program/robot.

Intentional Stance Through Expected Utility Maximization

The main issue with the intentional stance comes from a lack of operationalization; hence an easy tweak is to propose such a formalization. For example, Ronny's Comment on Coherence arguments do not imply goal directed behavior raised the idea that maybe a goal-directed system is one that is most efficiently described in terms of utility maximization. It's not clear from the post if Ronny knew about Dennett's work, but his proposal can still be interpreted as an operationalization of the intentional stance.

The main consequence is that goals are now well-defined as utility functions. But isn't that in contradiction with the test for goals? No, because of the constraint that the explanation be simpler than the mechanistic one. In Shah's post, the vacuity of utility maximization is shown by encoding the whole behavior in the utility function; this probably creates an explanation that is just as complex as a mechanistic one. Hence there is an implicit simplicity assumption for the utility function in Ronny's proposal. Nonetheless, how to capture this simplicity is still an open question.

Experimentally, Orseau et al.'s Agents and Devices: a Relative Definition of Agency use this operationalization for their agents, and they are able to infer goal-directed behavior in toy examples.

Hence this is one of the most promising concrete proposals for goal-directedness.

Richard Ngo's proposal

We've mentioned Ngo's AGI safety from first principles: Goals and Agency multiple times, when defining the tests for goal-directedness. In this post, Ngo proposes six criteria for goal-directedness:

1. *Self-awareness*: it understands that it's a part of the world, and that its behaviour impacts the world;
2. *Planning*: it considers a wide range of possible sequences of behaviours (let's call them "plans"), including long plans;
3. *Consequentialism*: it decides which of those plans is best by considering the value of the outcomes that they produce;
4. *Scale*: its choice is sensitive to the effects of plans over large distances and long time horizons;
5. *Coherence*: it is internally unified towards implementing the single plan it judges to be best;
6. *Flexibility*: it is able to adapt its plans flexibly as circumstances change, rather than just continuing the same patterns of behaviour.

Note that none of these traits should be interpreted as binary; rather, each one defines a different spectrum of possibilities.

How do these fare against our tests from the broader literature?

- **(What Goals Are)** As mentioned earlier, Ngo presents goals as made from the internal concepts of the system, rather than world state. This points towards a concrete space of goals, but is hard to evaluate for the moment as we don't have a good understanding (or even formalization) of the sort of concepts formed by a system.
- **(Explainability)** Since the criteria 2 (Planning) and 3 (Consequentialism) boil down to internal search, this proposal can use a mechanistic form of predictability hand in hand with the more intentional approach. It's thus satisfying for explainability.
- **(Generalization)** Generalization is a big theme in Ngo's AGI safety from first principles: Goals and Agency and all the AGI safety from first principles sequence. Assuming competence, a goal-directed system by Ngo's definition has all the ingredients needed to generalize: it looks for plans and adapts them to changes in the environment.
- **(Far-sighted)** Criterion 4 (Scale) explicitly requires that goal-directed systems care about the long-term consequences of their actions.
- **(Link with Competence)** This approach being based on internal structure, it doesn't require a minimal level of reachability to work. And efficiency should follow from the combination of many criteria like 2 (Planning), 3 (Consequentialism) and 6 (Flexibility).

Really, the only issue for our purposes with this definition is that it focuses on how goal-directedness emerges, instead of what it entails for a system. Hence it gives less of a handle to predict the behavior of a system than Dennett's intentional stance for example.

One surprising take is the requirement of self-awareness. In our readings, this is the only resource that conditioned goal-directedness on self-awareness (usually it's more related to agency, which makes sense because Ngo uses the two terms interchangeably). Is it a necessary component? Further research is needed to decide. But as it is far from a consensus in the literature, we don't include it in our list of tests.

Goal-Directed Intelligence

In contrast with the rest of this section, Vanessa Kosoy's proposal for goal-directedness is completely formal.

> Given $g > 0$, we define that "π has (unbounded) goal-directed intelligence (at least) g" when there is a prior ζ and utility function U s.t. for any policy π', if $E_{\zeta\pi'}[U] \geq E_{\zeta\pi}[U]$ then $K(\pi') \geq D_{KL}(\zeta_0||\zeta) + K(U) + g$. Here, ζ_0 is the Solomonoff prior and K is Kolmogorov complexity. When $g = +\infty$ (i.e. no computable policy can match the expected utility of π; in particular, this implies π is optimal since any policy can be *approximated* by a computable policy), we say that π is "perfectly (unbounded) goal-directed".

Let's ignore some details for the moment, and just consider the form of this definition: $\exists A, \forall \pi' : B \implies C$.

- A is a pair of a prior and a utility function.
- B is an inequality, asking that the expected utility parameterized by A of π' is greater than the one of π.
- C is another inequality, asking that the description complexity of π' be greater than the description complexity of the prior and the utility by at least g.

So in essence, having goal-directed intelligence g means that according to some goal, beating the policy costs at least g in terms of descriptive complexity.

It's not obvious what the link is to our tests, so let's spell it out.

- **(What Goals Are)** Kosoy appears to place herself in the expected utility maximization framework: goals are pairs of (utility function, prior).
- **(Explainability)** First, if goal-directed intelligence is infinite, the policy is optimal. This gives some form of explainability, in that the policy will accomplish its goal as well as possible.
  But what about policies with finite goal-directed intelligence? Intuitively, a policy easy to explain as optimizing the goal should have a low descriptive complexity. Yet the only constraint on descriptive complexity is a lower bound. If policy π has

goal-directed intelligence g, this tells us that the descriptive complexity of π is lower bounded by the descriptive complexity of the goal (prior and utility function) + g. So nothing forbids the descriptive complexity of π to grow with goal-directed intelligence, even if one would expect it not to.

- **(Generalization)** The existential quantifier in the definition captures the idea that there is some goal for which the property holds. So a discussion of generalization requires fixing the goal. And in an online setting like this one, generalization means something like minimizing regret, which here translates to maximizing expected utility. This works for infinite goal-directed intelligence; see the discussion of competence to see why this also works for finite goal-directed intelligence.
- **(Far-sighted)** No mention of long-term consequences of actions in this definition.
- **(Link with Competence)** Here ideal accomplishment means reaching the optimal expected utility for the goal, and efficiency is about how far one reaches towards that optimality. The ideal accomplishment part is quite irrelevant, as if the prior makes the optimal expected utility for a goal very small, there is still some sense in which goal-directed intelligence can increase.
  Concerning efficiency, Kosoy provided us with a simple result that goes a long way: if two policies have the same goal (the same goal satisfies the existential constraint for their goal-directed intelligence) then $g_1 > g_2$ implies that π_1 has better expected utility than π_2.[^8] Hence efficiency does increase with goal-directed intelligence.

The main issue with this proposal is not formalization, but understanding: it's a very compressed definition that needs further exploration to be judged more fully. That being said, we can already see that it ties goal-directedness with efficiency in a really strong way, fitting completely with our test.

Conclusion: What Needs to Be Done?

In this literature review, we extracted five tests for goal-directedness -- about goals, explainability, generalization, far-sightedness and competence. We also investigated how goal-directedness differed from optimization and agency, the two closest concepts studied in the AI Alignment literature. And we finally pitted various proposals for goal-directedness against the tests.

The conclusion is that two proposals seem particularly promising: the operationalization of the intentional stance through expected utility maximization, and Vanessa Kosoy's goal-directed intelligence.

That being said, there is still room for a lot of work in this sphere: questioning the criteria, proposing new definitions, and finding new intuitions to challenge the current proposals.

Our previous sequence, Toying With Goal-Directedness, proposed ideas in a less polished fashion, and already contains some of our thoughts on these questions. This sequence will expand on these ideas, and propose a structured and more thought through take on goal-directedness.

Notes

[^1] Simplicity is related to the idea of compression, which is the subject of one of our previous posts.

[^2] A slightly different take comes from Vanessa Kosoy's instrumental reward functions. Intuitively, such a function takes as input not a bare state, but an instrumental state -- an abstraction capturing the distribution of histories starting from this state according to the POMDP. This difference means that two states with different rewards are necessarily distinguishable by some policy (playing the role of an experiment); and thus that difference in rewards are indeed meaningful. We only mention this in a footnote as it's not clear how instrumental reward functions evade the issue pointed out by Shah.

[^3] This distinction might be related to the concept of locality we talked about in a previous post, and which will be studied further in this sequence.

[^4] Note that we don't mean complete generalization here. Only that in many cases the system will adapt its behavior to the environment.

[^5] Note that ideal accomplishment is related to generalization, in the sense that a high ideal accomplishment probably requires significant adaptation to the environment, and thus generalization. That being said, we are aiming here at something different from generalization: not just the ability to adapt, but the ability to "win".

[^6] This relates to the idea of focus that we presented in a previous post.

[^7] This part also applies to similar proposals that focus on explainability, for example the one by Max Tegmark.

[^8] The proof from Vanessa: since g_1 is the intelligence of π_1, for any policy π, if

$E[U](pi) \geq EU(\pi_1)$ then $K(\pi) \geq g_1 + C$. Since the intelligence of π_2 is $< g_1$ by

hypothesis, there exists π s.t. $EU(\pi) \geq EU(\pi_2)$ but $K(\pi) < g1 + C$. Applying the former

observation to the latter π, we get $EU(\pi) < EU(\pi_1)$. Combining the two inequalities

$EU(\pi_2) \leq EU(\pi) < EU(\pi_1)$. QED

Applications for Deconfusing Goal-Directedness

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Atonement for my sins towards deconfusion

I have argued that the deconfusion of goal-directedness should not follow what I dubbed “the backward approach”, that is starting from the applications for the concept and reverse-engineering its coherent existence (or contradiction) from there. I have also argued that deconfusion should always start and center around applications.

In summary, I was wrong. About the former

If deconfusion indeed starts at the applications, what about my arguments against the backward approach to goal-directedness? I wrote

> The same can be said for all convergent instrumental subgoals in the paper, and as far as I know, every argument about AI risks using goal-directedness. In essence, the backward approach finds out that what is used in the argument is the fundamental property that the forward approach is trying to formalize. This means in turn that we should work on the deconfusion of goal-directedness from the philosophical intuitions instead of trying to focus on the arguments for AI risks, because these arguments depend completely on the intuitions themselves.

My best answer is this post: an exploration of applications of deconfusing goal-directedness, and how they actually inform and constrain the deconfusion itself. The gist is that in discarding a different approach than what felt natural to me, I failed to notice all the ways in which applications do constraint and direct deconfusion. In this specific case, the most fruitful and important applications I’ve found are convergent subgoals, i replacing optimal policies, formalizing inner alignment and separating approval-directed systems from pure maximizers.

Thanks to John S. Wentworth for pushing hard on the importance of starting at the applications.

Applications

Convergent subgoals

Convergent subgoals (self-preservation, resource acquisitions…) are often important ingredients in scenarios starting with misspecified objectives and ending with catastrophic consequences. Without them, even an AGI would let itself be shut down,

greatly reducing the related risks. Convergent subgoals are also clearly linked with goal-directedness, since the original argument proposes that most goals lead to them.

As an application for deconfusion, what does this entail? Well, goal-directed systems should be the whole class of systems that could have convergent subgoals. It doesn't necessarily mean that most goal-directed systems will actually have such convergent subgoals, but a low-goal-directed system shouldn't have them at all. **Hence high goal-directedness should be a necessary (but not necessarily sufficient) condition for having convergent subgoals.**

This constraint then point to concrete approaches for deconfusing goal-directedness that I'm currently pursuing:

- Search for informal necessary conditions to each convergent subgoal, and then try to see the links/common denominator. Here I am looking for requirements which are simpler than goal-directedness, because they will hopefully be components of it.
- List for each convergent subgoals examples of systems with and without this goal, and search for commonalities.
- Based on Alex's deconfusion of power-seeking, look for a necessary condition **on the policies** for his theorems to hold.

Replacing optimal policies

When we talk about AGI having goals, we have a tendency to use optimal policies as a stand-in. These policies do have a lot of nice properties: they are maximally constrained by the goal, allow some reverse-engineering of goals without thinking about error models, and make it easy to predict what happens in the long-term -- optimality.

Yet as Richard points out in this comment, true optimal policies for real world tasks are probably incredibly complex and intractable. It's fair to say that for any task we cannot just enumerate on, we probably haven't built an optimal policy. For example AlphaGo and its successors are very good at Go, but they are not strictly speaking optimal.

The above point wouldn't matter if optimal policies pretty much behaved just like merely competent ones. But that's not the case in general: usually the very optimal strategy is something incredibly subtle that uses many tricks and details that we have no way of finding except through exhaustive search. Notably, the reason we expect quantilizers to be less catastrophic than pure maximizers is indeed that difference between optimal behavior and competent one.

Because of this, focusing on optimal policies when thinking about goal-directedness might have two dangerous effects:

- If the problems we investigate only appear for optimal policies, then it is probably a waste of time to study them, as we won't build an optimal policy (or not for a very long time). And wasting resources might prove very bad for shorter timelines.
- If the problems we investigate also appear for merely competent goal-directed policies, but we have to wait for optimality before spotting them when

training/studying something, we're fucked because they will crop up way before that point.

What I take from this analysis is that **we want to replace the optimality assumption by goal-directedness + some competence assumption.**

Here we don't really have a necessary condition, so unraveling what the constraint entails is slightly more involved. But we can still look at the problems with optimality, and turn them into requirements for goal-directedness:

- Since optimal policies don't capture the competent policies we're actually building, we want goal-directedness to do it.
    - Possible approach: list the competent AI we're able to produce, and try to find some commonality beyond being good at their task.
- But not all policies should be goal-directed, or it becomes a useless category.
    - Possible approach: find examples of policies which we don't want to include in the goal-directed ones, possibly because there is no way for them to get convergent subgoals.
- Less sure, but I see the issues with optimality as stemming from an obsession with competence. This confort me in my early impression that goal-directedness is more about "really trying to accomplish the goal" than in accomplishing it.
    - Find the commonality between not very competent goal-directed policies (like an average chess player), and try to formalize it.

Grounding inner alignment

Risks from Learned Optimisation introduced the concept of mesa-optimizers or inner optimizers to point to the results of search that might themselves be doing internal search/optimization. This has been consistently confusing, and people constantly complain about it. Abram has a recent post that looks at different ways of formalizing it.

In addition to the confusion, I believe that focusing on inner optimizers as currently defined underestimate the range of models that would be problematic to build, because I expect some goal-directed systems to not use optimization in this way or have explicit goals. I also expect goal-directedness to be easier to define than inner optimization, even if that probably comes from a bias.

Rephrasing, the application is that **goal-directedness should be a sufficient condition for the arguments in Risks to apply**.

The implications are quite obvious:

- Mesa-optimizers should be goal-directed
    - Possible approach: look for the components of a mesa-optimizers, and see what can be relaxed or abstracted while still keeping the whole intuitively goal-directed.
- Goal-directed systems should have the same problems/issues argued in Risks.
    - Possible approach: find necessary conditions for the arguments in Risks to hold.
- Goal-directedness should be less confusing than inner-optimization/mesa-optimization.

    - Possible approach: list all the issues and confusions people have with inner optimization, and turn that into constraints for a less confusing alternative.

Approval-directed systems as less goal-directed

This application is definitely more niche than the others, but it seems quite important to me. Both Paul in his approval-directed post and Rohin in this comment to one of his posts on goal-directedness have proposed that approval-directed systems are inherently less goal-directed than pure maximizers.

Why? Because approval-directed systems would have a more flexible goal, and also wouldn't not have the same convergent subgoals that we expect from competent goal-directed systems.

I share this intuition, but I haven't been able to find a way to actually articulate it convincingly. Hence why I add this constraint: **approval-directed systems should have low-goal-directedness (or at least lower than pure-maximizers)**

Since the constraint is quite obvious, let's focus on the approaches to goal-directedness this suggests.

- Deconfuse approval-directed systems as much as possible, to have a better idea of what their low goal-directedness would entail
- List all the intuitive differences between approval directed systems and highly goal-directed systems
- Look for sufficient conditions (on a definition of goal-directedness) for approval-directed systems to have low goal-directedness.

Conclusion: all that is left is work

In refusing to focus on the application, I slowed myself down in two ways:

- By going into weird tangents and nerd-snipe without a mean to check if the digression was relevant or not
- By missing out on the many approaches and research directions that emerge after even a cursory exploration of these applications.

I attempted to correct my mistake in this post, by looking at the most important applications for deconfusing goal-directedness (convergent subgoals, optimality, inner optimization and approval directedness), and extracting constraints and questions to investigate from them.

This cuts my work for me on the topic; if you find yourself interested or excited by any of the research ideas proposed in this post, send me a message so we can talk!

Goal-Directedness and Behavior, Redux

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Beyond past confusions

Over the last year, I wrote and thought many confused and confusing ideas on the relationship between goal-directedness of behavior. In the linked post for example, I defended a deconfusion of goal-directdness solely in terms of behavior; in doing so, I might pass for a behaviorist (someone thinking that mental constructs are not needed and so don't exist), or look like I imply that we should never use internal knowledge of our models to determine goal-directedness. Without even mentioning the factual errors.

So here is my attempt at a short and clear explanation of the link I see between goal-directedness and behavior. If you're confused by this take, or believe me to be confused, I would really appreciate a comment. My goal isn't to prove that I'm obviously right, just to get less confused and hopefully help lift the fog of confusion for everyone.

Thanks to Jack Koch for a recent discussion that reminded me of this issue, and to Richard Ngo for giving me food for thought on this subject with his comments .

Behavior in all its glory

What's my take? I think that when we talk about goal-directedness, **what we really care about is a range of possible behaviors**, some of which we worry about in the context of alignment and safety. We might for example think that goal-directed systems have convergent subgoals, which tells us how they could lack corrigibility and cause catastrophic outcomes. that such a goal-directed system could follow.

My entire point is that for deconfusing goal-directedness, we want a better understanding of this range of behaviors. At the moment, when thinking about a given behavior, I don't know whether that's the sort of thing a goal-directed system would do. And it seems problematic both for understanding the risks of goal-directed systems, and for detecting them.

Note that even a purely structural definition of goal-directedness would constrain the structure such that the system behave in a certain way. So even if we want a structural definition, clarifying the range of behaviors sounds like progress.

What I'm not saying

- We shouldn't ascribe any cognition to the system, just find rules of association for its behavior (aka Behaviorism)

    - That's not even coherent with my favored approach to goal-directedness, the intentional stance. Dennett clearly ascribes beliefs and desires to beings and systems; his point is that the ascription is done based on the behavior and the circumstances.
- Nothing but the behavior is useful to check goal-directedness.
    - Even in my original confused post, I point out that structural knowledge about the system is probably necessary to check goal-directedness, as its probably the only tractable way of finding out what the system will do.
    - I hadn't thought about it last year, but I see more and more the value of thinking about the justified beliefs that the system might have, due to its training data, learning algorithm and inductive biases. (This is an idea of Paul with ties to universality)

How I could be wrong

The main crux I see about this take on behavior is whether it's even possible or tractable to deconfuse and formalize the range of behaviors of goal-directed systems. No matter how useful a formalization would be, if we can't get it, we should turn to other approaches.

That being said, I haven't seen any convincing argument that it's impossible, and the more I dig, the more stuff I find, so I am quite convinced that some progress is possible.