Desiderata for a mathematical theory of decision-making

Tsvi Benson-Tilsen

This is a collection of my current hypotheses about what decision theory is for and what would constitute progress in decision theory.

***

There are not many nonlocal dependencies, so feel free to skip around liberally. I've chosen not to track sources, both for my own convenience and also because I may have put my own spin on the ideas; but it is safe to assume that the ideas and my apprehension of them are due almost entirely to conversations with and writings by other reasearchers. This should be considered a draft, so comments are particularly welcome.

1 What decision theory is for

Decision theory is for

- directing a consequentialist reasoning process to reason about how to achieve particular goals; and
- directing autonomous agents to reliably achieve particular goals.

"Directing a reasoning process" is not restricted to decision-making systems. Broadly, if we are getting useful work out of an AI, it's because the AI is doing computations that produce logical information about how we can achieve our goals. So we must have somehow pointed the consequentialist-reasoning part of the AI towards our goals. The questions of decision theory in this context are then: How do you specify logical or environmental goals? How do you direct a consequentialist reasoner to produce the relevant logical information? How do you make a consequentialist reasoner that does this efficiently?

There may be more specific problems relating to agent-like systems in particular. For example, systems that make decisions may exhibit reward hacking, instability under self-modification, averting instrumental incentives, severe suboptimality due to things like spurious counterfactuals, and other more exotic failure modes such as Pascal's mugging. A good theory of decision-making might also enlighten us about other research paths such as low impact, mild optimization, and so on.

I think it's plausible that it's simply not necessary to have a deeper theoretical understanding of decision-making in order to design AIs that are capable and aligned enough to fulfill their potential for positive impact. For example, from this perspective, Paul's approach could be viewed as saying that decision theory problems, especially the more philosophical ones, can be solved satisfactorily by appealing to and amplifying human judgement. Even if a mathematical theory of decision-making were necessary, this is a fact we might want to first learn more definitively and crisply, and then solve. However, decision theory does seem like an identifiable fundamental theoretical aspect of AI design that may take long serial time to solve, so I think it's still very worth pursuing.

2 Desiderata for decision theory

This is a list of mostly informal, possibly overlapping desiderata for a decision procedure or for a general theory of decision procedures. Note that the categorization is not clean; for example, understanding logical counterfactuals cuts across all categories. Further, some of these problems impinge on other parts of AI alignment theory; for example, reward hacking might also be viewed as a value specification / reward engineering problem. These problems don't all carry the same philosophical importance or the same practical importance.

It seems worthwhile to formalize desiderata (these or others) for decision theory. In my current view, a formalization or a satisfaction of these desiderata is judged successful or not according to how enlightening it is about the two bullets in the previous section.

Many of these desiderata are parametrized by considering them in different settings and with different constraints. For example, agents can be judged according to:

- their performance given different resources such as memory, space, time, information, randomness, oracle access, proof-theoretic strength, etc.;
- their performance compared to the performance of different classes of other agents or other policies, or their performance under some other measure such as total expected utility, local optimality under some system of counterfactuals, or other forms of regret;
- their performance in the limit, asymptotically, or practically, along different parameters; and
- their performance in different environments such as ones that are cartesian, are more or less computationally or information-theoretically complicated, contain other agents, are Markovian or otherwise constrained, and so on.

2.1 Optimal bounded reasoning

Roughly, a bounded reasoner R is "good" if it answers as many questions, as accurately, as inexpensively, as possible. We'll know we have a solution when we can prove theorems giving optimal reasoners or impossibility results for various formal models of the problem class and optimality criterion. For example:

- **R should be theoretically efficient and have accurate beliefs.** That is, as a precursor to practical efficiency, R should have good computational complexity properties, possibly using realistic theoretical assumptions about the sorts of problems we will use R for. More precisely, we can ask for "no wasted motion": if there is some way to reliably answer some question correctly in a certain amount of time, then R doesn't spend much more time than that.

- R **should be practically efficient and correct.** That is, eventually, we want to actually run R usefully in real life and have the theoretical results still apply.

2.2 Philosophically sound reasoning

A good theory of reasoning should make us unconfused about bounded, embedded reasoning. For example:

- R **can refer to the environment.** That is, parts of R's cognition reliably correspond to parts of the world, so that assertions made by R can be reasonably interpreted (either by a human or in principle) as assertions about the world.

  This is already solved to some extent by existing AI frameworks, but still seems to be a confusion that crops up in many highly-capable agent designs.

- R **reasons naturalistically.** That is, R reasons and processes observational evidence as though it is a part of the world it is reasoning about, rather than an ephemeral observer. For example, R should handle anthropic reasoning sanely, and should be self-locating in the sense that it uses its observations to eliminate its "indexical uncertainty".

  This may be achieved in a logical setting already by conditioning on statements of the form "my algorithm gets such-and-such input".

- R **has good third-person counterfactuals.** That is, we can ask R questions of the form "What if we (humans) implement such and such plan?", and get an answer that we would intuitively consider to capture the meaning of the question.

  This seems like a thorny philosophical problem, but also seems quite important.

- R **can reason sanely about exotic environments.** That is, R does something reasonable in the face of environments that e.g. contain copies of R or reasoners similar to R, permit various forms of hypercomputation or time travel, are otherwise uncomputable, are otherwise outside the ontology natively used by R, and so on; as opposed to refusing to believe the evidence, inventing complicated implausible explanations for the evidence, etc.

  This seems like a problem we should be able to defer to future versions, and doesn't obviously promise much in the way of safety-relevant understanding.

2.3 Optimal decision-making

Roughly, a decision-making procedure D is "good" if an agent A that makes decisions using D gets highly-valued outcomes in as many settings as possible. We'll know we have a solution when we can prove theorems giving optimal agents or impossibility results for various formal models of the agent, problem class, and optimality criterion. For example:

- A **should take optimal actions in fair problems.** That is, for any given environment, if success of any agent in that environment depends only on that agent's actions, then A should do as well as any other policy does in that environment.

  In Cartesian environments such as standard reinforcement learning environments, which are generally fair, a number of algorithms such as Sarsa find optimal policies, though perhaps not optimally in terms of convergence rates or computational efficiency. More general RL environments are more complicated; compare AIξ. Optimality can be achieved to some extent by modal agents in the proof-based setting, and by agents with access to a Garrabrant inductor in the asymptotic decision theory setting.

- A **should perform optimally in resource-bounded fair problems.** That is, if an environment is fair except that it penalizes A based on its usage of computational resources like time and space, then A should receive as high utility as other agents. For example, A should answer computational decision problems as quickly as is optimal, and A should succeed in learning, optimization, and game playing tasks.

  This is mostly open, though asymptotic decision theory is relevant. I plan to work on this, as it seems tractable; optimality notions can be imported from the abovementioned fields, and then satisfied individually and then simultaneously by one agent.

- A **should make the correct tradeoffs between performance in different environments.** That is, if e.g. there is uncertainty about the environment, then actions should be taken to do well in environments that are more likely or more important or more amenable to being influenced.

  This is generally a simple consequence of the definition of an expected utility maximizer, but may be less immediate for other agent concepts such as reinforcement learners, agents with alternative expected utility targets like satisficers or quantilizers, and policy ensemble ("masquerade") algorithms. This is also relevant to problems in anthropics.

- A **should avoid spurious counterfactuals.** That is, there shouldn't be a "lock-in of beliefs about conterfactuals", as in: I think that if I take the \$10 bill, then bees, so I never take the \$10 bill, and therefore don't find out that doing so would not in fact cause bees.

  Reinforcement learning generally avoids this problem by exploring, and thereby putting beliefs about counterfactuals to the empirical test. More straightforward EU maximizers often suffer from spurious counterfactuals, e.g. by Lob's theorem in proof-based agents and by "lock-in" of strongly-held false beliefs for agents that reason by induction.

- *A* **should not explore dangerously.** That is, *A* should not take actions that may have catastrophic effects, purely in the name of gathering information.

  This seems in direct conflict with avoiding spurious counterfactuals. We could argue that VOI considerations should always produce a good answer to the exploration-exploitation trade-off. One issue is that the VOI calculations may themselves be based on fundamentally bad reasoning, which can only be corrected by exploring; and this fact may itself be difficult to understand using only the same faulty reasoning.

- *A* **should account for the fact that it is embedded.** That is, if the fact that *A*'s cognition takes place inside the universe that it is modeling is relevant to *A*'s decision making, then *A* should take that fact into account (by making decisions "naturalistically"). For example, if thinking too hard would cause *A*'s circuits to overheat and fail, then *A* shouldn't think too hard.

  To my knowledge there is no general theory of decision-making that accounts for this. Failures of this form should be corrected by reinforcement learning, insofar as the failures can be modeled within the training environment. However, in other cases RL doesn't seem to suffice. For example, suppose *A* can take actions that lead to a kernel panic that interrupts the training process. These actions won't get negatively reinforced, and so there is a selection bias in *A*'s training data, where sufficiently bad outcomes aren't accounted for at all. Hence this problem seems to reflect a hole in our theory, and failures such as wireheading can be viewed as a failure of embedded reasoning.

- *A* **should perform computations according to decision-relevance.** That is, *A* should choose which computations to run depending on their cost weighed against their "value of (logical) information", e.g. the expected increase in utility of *A* acting with the knowledge of that computation as opposed to without it.

  This is essentially an open problem as far as I know, and would appreciate pointers to relevant literature. This might be a special case of naturalistic decision-making.

- *A* **should handle logical uncertainty about the world without biasing its decisions.** For example, one could imagine that *A*'s decisions might be overly determined by considerations in possible worlds that are difficult to prove impossible, e.g. because there are always worlds that are plausible and promise huge returns, but are not actually possible. Pascal's mugging may be an instance of this, though I'm unsure.

  This may be a non-issue—Garrabrant inductors might handle this issue with no problem.

- A **should handle logical uncertainty about its utility function.** For example, implementing indirect normativity may require that the utility function is very difficult to compute. In the same vein, counterfactual oversight seems to demand a hard-to-evaluate reward function. The expected value of an action depends on the uncertainty about the utility function, as well as about the outcomes.

- A **should handle exotic environments.** That is, A should do something sensible, even when placed in environments that might be considered pathological. For example, A should make appropriate use of halting oracles, time travel, infinite computational resources, and so on.

  This desideratum doesn't seem very important.

- A **should account for logical dependencies in the environment.** That is, roughly speaking, A should act as though it has decision-control over the parts of the environment, such as other agents, that instantiate or reason about A. For example, A should pay up in the counterfactual mugging, select one box in (transparent) Newcomb's problem and in agent simulates predictor, and press the button in the procrastination paradox.

  I consider satisfying this desideratum to be a key open problem in decision theory. Timeless decision theory and updateless decision theory are significant progress, but both are underspecified. Both require a notion of counterfactuals on logical facts, and while UDT fixes some problems with TDT, it is not clear how to be updateless with respect to logical facts.

  The remaining desiderata—coordination, cooperation, and invulnerability to adversaries—can be viewed as special cases of logical dependencies.

- A **should coordinate with itself.** That is, different copies of A should be able to coordinate to execute plans that rely on multiple actions across space and time. For example, A should meet up at one of two coffee shops with its copy without needing to communicate, and make plans that its future self can and will execute.

  This is theoretically trivial in cases where UDT can be applied directly. On the other hand, it is not trivial to coordinate across time (that is, across instances of A with different computational power). In such cases, even an implementation of UDT that is updateless with respect to logical facts has the problem that it is supposed to coordinate with earlier versions of itself—but those versions had worse opinions about what policies are good, and so it is not clear when and how to actually follow through with whatever precommitments the earlier versions may have made.

  This desideratum may be a special case of cooperating with cooperative agents.

- A **should cooperate with cooperative agents.** That is, A should successfully execute plans that rely on both A and other agents following a certain joint policy. For example, A should cooperate in the Prisoner's dilemma with agents symmetric to A, and A should achieve Pareto-optimal outcomes (or at least better than disagreement-point outcomes) in bargaining dilemmas.

I haven't looked into the literature on bargaining, and would like to hear about the relevant work. Particularly interesting is bargaining between agents with different information / different computational resources.

- **A should not be vulnerable to adversarial parts of the environment.** That is, A should be harmed as little as possible by parts of the environments (e.g. other agents) that model A and execute plans to interfere with A. For example, A probably ought to ignore blackmail (causal, evidential, or otherwise) and should avoid being exploited in bargaining dilemmas, and A should exploit other agents that can be exploited.

  This may just be a special case of accounting for logical dependencies, or it may present particular new challenges.

- **A should not be vulnerable to adversarial parts of its cognition.** That is, A should avoid cases where some subroutine or substructure of its cognition acts at cross-purposes to A's goals (from within A). In more words, suppose A is made up of subagents, or uses black-box or ensemble methods to model or predict the world. It may be that some of those subroutines are themselves agents ("adversarial hypotheses" in the case of ensemble methods). Then those agents should not be able to seize control of A or otherwise impact the world negatively by influencing A's actions, even if those agents can coordinate with each other. For example, see steganography and simulation warfare.

  This might be solvable by the following: not having adversarial hypotheses in the first place, e.g. by only running computations that have been whitelisted by some explicit decision-making process; very strong "containment" of subroutines, i.e. proving that A's decision procedure can't possibly be influenced negatively by subroutines; or as a special case of naturalistic decision making.

- **A should not be vulnerable to adversarial parts of its utility function.** I list this as a separate special case of adversarial parts of cognition because proposals such as indirect normativity potentially have this problem, and it might slip under the radar since "whatever the utility function says is good, must be good". In that respect it is more a question of value specification, but has a decision-theoretic flavor.

2.4 Philosophically sound decision-making

A good theory of decision-making should make us unconfused about rational bounded agency. We'll know we have a solution when we can analyze agent designs and give persuasive and empirically accurate accounts of whether they are taking actions for reasons we would endorse. For example:

- *A* **should be reflectively stable.** That is, *A* should endorse its current and future beliefs and actions, such that it doesn't have an incentive to modify the way it selects those beliefs and actions. For example, stability might fail for agents that: perform suboptimally in fair problems, since they could do better by changing that behavior; two-box in Newcomb's problem, since they could do better by precommitting to one-box; or refuse to pay in the counterfactual mugging, since before knowing the outcome of the coin, they have a higher expected value for precommitting to pay up.

  This seems like a crucial desideratum. If we want to analyze an agent usefully and have that analysis stay relevant over time, we probably want the assumptions we used about the agent to continue holding. It seems not too hard to make an agent that is technically reflectively stable; what is more interesting is to find a "natural" or "interpretable" decision procedure that is also reflectively stable. It is unclear what self-modifications we should endorse (such as removing biases, making implementations faster or more robust, and so on), but a reasonable first goal is having an agent that is reflectively stable at all.

- *A* **should make decisions naturalistically.** That is, *A* should learn the effects of its actions on the environment without having to draw any false distinctions between its cognition and the environment.

  For example, *A* shouldn't assume a cartesian boundary between itself and the world, and hence should appropriately account for effects of the environment on its cognition (such as sped up, slowed down, or otherwise distorted cognition caused by modifications to *A*'s implementation), or vice versa (such as heat caused by information processing used in *A*'s implementation). Also, *A* should should select which cognition to perform automatically as an instrumental goal of its objective.

- *A* **should not interfere with the physical implementation of its utility function.** That is, *A* should not wirehead. I think that generally, a goal-based agent with environmental goals won't wirehead, since altering its utility function will make it less effective at pursuing its current utility function; but other frameworks such as reinforcement learning are liable to wirehead.

  I classify this as a philosophical problem because it is unclear to me how to say, philosophically, that a wireheading agent is doing something that is not "requested" by a utility or reward function that can be satisfied by wireheading; but of course we don't want agents to wirehead.

- *A* **should not exert perverse logical influence on its utility function.** That is, *A* should not "logically" wirehead, i.e. make decisions in order to alter the result of computations that define *A*'s utility function applied to some outcome. For example, proposals for indirect normativity generally involve predicting the output of a very-long running computation that involves humans as well as (potentially) *A* or analogs of *A*; this gives

A the opportunity to affect the outcome of utility judgements, e.g. to give utility judgements that are very easy to score well on.

I know of no direct work on this, except possibly old work by Paul. This may be more a question of goal specification, but decision theory might provide relevant insights. I haven't worked this through in detail, but I suspect that a number of AI frameworks, at least in the theoretical realm of extremely high capabilities, would fall prey to this; for example, I (weakly) conjecture that Everitt and Hutter's VRL agent would manipulate a sufficiently accurate learned utility function.

- **A should have accurate beliefs about first-persion counterfactuals.** That is, A's model of "the world in which I take this action given this input" is "as correct as possible".

  Having "correct counterfactuals" seems like a very abstract desideratum—of a philosophical nature, hard to formalize, hard to make progress on, hard to know when you're done—but also potentially very important, as many other problems might reduce to counterfactuals. For example, "taking into account logical dependencies", "naturalistic reasoning", and bounded optimality might be solved by a good theory of counterfactuals.

- **A should be able to pursue exotic preferences.** For example, A should be able to represent and optimize for preferences that evaluate or depend on preferences held by other agents, preferences about computations or logical facts, preferences with non-physical referents such as how some event is caused, preferences about how A forms beliefs or makes decisions, preferences that are difficult to compute, unbounded utility functions, and so on.

- **The theory should clarify the relationship between utility, belief, and reality.** That is, we should become unconfused about the distinctions and relationships between preferences, beliefs about the world, and the world itself. For example: anthropic reasoning, indexical uncertainty, a "caring measure" over logically consistent worlds, a "caring measure" over possibly inconsistent worlds, and computationally difficult utility functions.