Counterfactual Reprogramming Decision Theory

Ben Goertzel
Novamente LLC
1405 Bernerd Place
Rockville MD 20851

April 25, 2010

Abstract

A novel variant of decision theory is presented. The basic idea is that one should ask, at each point in time: *What would I do if the reprogrammable parts of my brain were reprogrammed by a superintelligent Master Programmer with the goal of supplying me with a program that would maximize my utility averaged over possible worlds?* Problems such as the Prisoner's Dilemma, the value of voting, Newcomb's Problem and the Psychopath Button are reviewed from this perspective and shown to be addressed in a satisfactory way.

1 Introduction

Decision theory has proved a surprisingly thorny topic, with a number of approaches proposed and none considered fully satisfactory. Evidential decision theory (EDT) [GH81], the most straightforward and traditional approach, gives problematically counterintuitive results in a number of situations [Lew81], and so the field has shifted attention to a variant called causal decision theory (CDT) [Lew81], which solves some of these issues. Conceptually, causal decision theory isolates certain actions as causally unrelated to certain other events; there are various ways of formalizing this isolation, but none is yet widely accepted. And there are some examples on which causal decision theory also gives counterintuitive results [Ega07], as well as examples on which its recommendations seem less ethical, and no more rational, than those of EDT.

Here we present an alternate approach, which we call Counterfactual Reprogramming Decision Theory (CRDT), which appears to avoid the problems with both EDT and CDT. The basic idea of CRDT is to ask, at each point in time, what one would do *if* the reprogrammable parts of one's brain were reprogrammed by a superintelligent Master Programmer with the goal of supplying one with a program that would maximize one's utility averaged over possible worlds. Then, one carries out this action.[1]

After outlining the CRDT idea in more detail, we review a series of decision-theoretic problems commonly considered troublesome: the Prisoner's Dilemma, related issues regarding voting, Newcomb's Problem and Nesov's variation, and the Psychopath Button. In each case we explain how CRDT handles the problem unproblematically.

The treatment here is only semi-formal; mathematical notation is used, and quantitative calculations are done, but there are no theorems and proofs. The Appendix sketches a formalization of CRDT in terms of a formal agents model, but a full formalization of CRDT is left for another paper.

2 Evidential and Causal Decision Theory

Contemporary decision theory involves the application of probability theory to estimate the utility of various possible courses of action. The underlying mathematics is straightforward, setting aside issues of computa-

[1] Some interesting, albeit often confusing, discussion on CDT and hypothetical replacement decision theories may be found online at the *Less Wrong* blog [Dea09]. The decision algorithm presented by Dai on that blog page bears some resemblance to CRDT, but due to the rough and very informal exposition there, I'm not sure what is the precise relationship.

tional complexity in complex situations. But the application and interpretation of the relevant formulas can be subtle.

The classical approach is evidential decision theory, according to which the best action is the one which, conditional on your having chosen it, gives you the best expectations for the outcome. But this gives rise to some strange recommendations in cases where there are common historical factors underlying one's actions and the events in one's environment. To work around these issues, causal decision theory has been introduced: it requires a causal connection between your actions and the desirable outcome. However, the foundations of causal decision theory are a thorny topic and the focus of continual debate in the philosophy literature [Ega07]; and the approach also runs into trouble on some simple examples.

2.1 Evidential Decision Theory

Consider an agent Z deciding what action to take in a certain situation, and assume the action takes place during a time interval T.

Let

- $\{U_1, ..., U_n\}$ = a partition of the set of states of the universe, which may be dependent on agent Z
- O_i = the proposition that the universe has a state in partition element U_i during time-interval T
- A = the event that agent Z does an action in class Ac during time interval T
- $U(A \cap O_i)$ = expected benefit if Z does Ac during T and universe is in state-set U_i during T = average utility-to-Z of all specific universe-states within U_i in which the agent does Ac
- $P(O_i)$ = probability that the universe will be in state-set U_i during interval T
- $P(A)$ = probability that Z takes an action in class Ac during interval T
- $P(X|A)$ denote conditional probability (the probability of X given A)

Then there is an elementary, familiar theorem telling us that the expected utility of the event A is

$$U(A) = U(A \cap O_1)P(O_1|A) + ... + U(A \cap O_n)P(O_n|A)$$

The subtle term in this formula is $P(O_i|A)$. Of course, since the agent Z has never before had the opportunity to carry out (or not) an action in class Ac during time interval T, this probability must necessarily be estimated based on indirect evidence. Making this sort of estimate is mathematically straightforward given appropriate evidence, but various conceptual issues arise – all pertaining to the case where there are common factors determining whether Z takes an action in Ac, and whether the universe is in state U_i.

2.2 Causal Decision Theory

To see the motivation for proposing causal decision theory as an alternative to evidential decision theory, consider the following example, presented by Gibbard and Harper in their original paper on CDT. Suppose that King Solomon wants Bathsheba but fears that summoning her would provoke a revolt. Further, suppose that Solomon

has studied works on psychology and political science which teach him the following: Kings have two personality types, charismatic and uncharismatic. A king's degree of charisma depends on his genetic make-up and early childhood experiences, and cannot be changed in adulthood. Now, charismatic kings tend to act justly and uncharismatic kings unjustly. Successful revolts against charismatic kings are rare, whereas successful revolts against uncharismatic kings are frequent. Unjust acts themselves, though, do not cause successful revolts; the reason uncharismatic kings are prone to successful revolts is that they have a sneaky, ignoble bearing. Solomon does not know whether or not he is charismatic; he does know that it is unjust to send for another man's wife. (p. 164)

EDT tells us that Solomon should abstain from Bathsheba. But this seems wrong, since whether Solomon is charismatic or uncharismatic cannot be changed. Intuitively, it seems Solomon should choose to send for her.

To encompass this intuition, causal decision theory defines the expected utility U of an action A as

$$U_{\text{causal}}(A) = U(A \cap O_1)P(O_1 < A) + ... + U(A \cap O_n)P(O_n < A)$$

where $P(A > O_j)$ is the *counterfactual probability* that, if A were done, then O_j would hold.

The appropriate definition of this "counterfactual probability" has proved a subtle and controversial matter. Gibbard and Harper showed that if we accept two simple axioms, then the statistical independence of the events A and $A > O_j$ suffices to yield $P(A > O_j) = P(O_j|A)$. However, there are cases in which actions and conditionals are not independent – and as noted above, these are precisely the situations where EDT give intuitively strange results. David Lewis [Lew81] showed that, under one plausible definition, the probability of a conditional $P(A > O_j)$ does not always equal the conditional probability $P(O_j|A)$.

In the Bathsheba example, we have four outcomes

- BR = have Bathsheba and get revolted against
- $\tilde{B}R$ = not have Bathsheba, but get revolted against
- $B\tilde{R}$ = have Bathsheba and not get revolved against
- $\tilde{B}\tilde{R}$ = not have Bathsheba and not get revolted against

and

$$U_{\text{causal}}(A) = U(BR)P(BR < A) + U(\tilde{B}R)P(\tilde{B}R < A) + U(B\tilde{R})P(B\tilde{R} < A) + U(\tilde{B}\tilde{R})P(\tilde{B}\tilde{R} < A)$$

The problem statement suggests that no action A Solomon can undertake will affect the choice of whether he gets revolted against or not, which means we can write

$$P(BR < A)/P(R) = P(B\tilde{R} < A)/P(\tilde{R}) = P(B < A)$$
$$P(\tilde{B}R < A)/P(R) = P(\tilde{B}\tilde{R} < A)/P(\tilde{R}) = P(\tilde{B} < A)$$

and hence

$$\begin{aligned} U_{\text{causal}}(A) = \\ U(BR)P(B < A) + U(\tilde{B}R)P(\tilde{B} < A) + U(B\tilde{R})P(B < A) + U(\tilde{B}\tilde{R})P(\tilde{B} < A) = \\ (U(BR) + U(B\tilde{R}))P(B < A) + (U(\tilde{B}\tilde{R}) + U(B\tilde{R}))P(\tilde{B} < A) \end{aligned}$$

But since

$$U(\tilde{B}R) < U(BR)$$
$$U(\tilde{B}\tilde{R}) < U(B\tilde{R})$$

it follows that Bathsheba should be summoned, i.e.

$$U_{\text{causal}}(\text{summon}) > U_{\text{causal}}(\text{not summon})$$

The key steps here of course are the reductions such as $P(BR < A) = P(B < A)P(R)$. It is *not* the case in this example that $P(BR|A) = P(B|A)P(R)$ and so forth (because of the dependency between A and R, considered from the perspective of statistical analysis over possible worlds), so this is a genuine difference from evidential decision theory.

3 Counterfactual Reprogramming Decision Theory

We suggest a fundamentally different approach. Instead of thinking on the level of *actions*, we suggest to think on the level of *programs*. At any given point T, we suggest, an agent may carry out the following thought-experiment:

- Assume that the agent's brain is partially reprogrammable, but also has certain *immutable properties*
- Imagine a Master Programmer (MP), able to replace the reprogrammable portion of the agent's brain with an arbitrary computer program of length $< L$ and runtime $< R$, where L and R are large numbers.
- The goal of the Master Programmer is to replace the reprogrammable portion of the agent's brain with a program $\mathcal{P}$ having the property that, averaged over all possible worlds that are consistent with the agent's current world-knowledge (and with a weighting in the average, so that possible worlds considered more likely by the agent are weighted higher), operating $\mathcal{P}$ will cause the agent to get maximal utility
- Imagine that the Master Programmer replaces the reprogrammable portion of the agent's brain with a new program $\mathcal{P}$ right now, at time T
- Figure out what action $\mathcal{P}$ would take, and then take that action

The theory that agents should follow this advice in choosing their actions, is what we call *Counterfactual Reprogramming Decision Theory* or CRDT. A more formal version of CRDT is presented in the Appendix, using a simple formal agents model drawn from [LH07a].

Note that the subtleties mentioned above regarding $P(O_i|A)$ are irrelevant here. Because, the Master Programmer is assumed to make his judgments by purely mathematical criteria, so there is no "common cause" between his choice of action and the outcome O_i.

Of course, this decision strategy may be difficult to implement in many cases, because most agents are not master self-programmers. However, the CRDT procedure described above at least serves as a "gold standard" for guiding decisions. And in many cases it's possible to psych out what actions the Master Programmer's program would recommend, without actually being a Master Programmer.

For example, consider the case of Solomon and Bathsheba. It's quite clear what the Master Programmer would do. If given the opportunity to reprogram Solomon's brain at the time point of the story, he would reprogram Solomon to summon Bathsheba. Because this will give Solomon the greatest utility, going forward. (We assume that the immutability of Solomon's charisma or otherwise is given as a "hard constraint" and wired into the *non-reprogrammable* portion of the agent's brain. So the Master Programmer isn't able to reprogram Solomon to be charismatic.) Thus, according to CRDT, Solomon should summon Bathsheba.

4 The Prisoner's Dilemma

Now we review the application of EDT, CDT and CRDT to the (one-shot) Prisoner's Dilemma. In this context, CRDT leads to an immediate and simple explanation of the phenomenon Douglas Hofstadter called "superrationality" [Hof79].

The Prisoner's Dilemma (PD) is typically articulated as [Wik10b]:

Two suspects are arrested by the police. The police have insufficient evidence for a conviction, and, having separated both prisoners, visit each of them to offer the same deal. If one testifies (defects from the other) for the prosecution against the other and the other remains silent (cooperates with the other), the betrayer goes free and the silent accomplice receives the full 10-year sentence. If both remain silent, both prisoners are sentenced to only six months in jail for a minor charge. If each betrays the other, each receives a five-year sentence. Each prisoner must choose to betray the other or to remain silent. Each one is assured that the other would not know about the betrayal before the end of the investigation. How should the prisoners act?

4.1 One-Shot Prisoner's Dilemma with Identical Players

Applying EDT to PD with identical players, for Player 1 we find:

$$U(C) = U(C \cap C, C)P(C, C|C) + U(C \cap C, D)P(C, D|C) = \text{ payoff}(C, C)$$
$$U(D) = U(D \cap D, D)P(D, D|D) + U(D \cap D, C)P(D, C|D) = \text{ payoff}(D, D)$$

So according to this, it is rational for Player 1 to cooperate.

CDT, on the other hand, tells a different story. In the CDT point of view, Player 1's actions have no causal impact on Player 2, during the course of the PD game. Thus, CDT advises Player 1 to defect. This is what Hofstadter calls basic rationality, as opposed to the "superrationality" suggested by EDT. The dichotomy between EDT and CDT here suggests that, in some cases at least, EDT is ethically superior to CDT. One may also observe that, in essence, CDT rejects the problem statement here. The problem statement tacitly assumes a "non-causal" connection between the two players, which CDT cannot encompass.

What does CRDT say? The "identical players" constraint may be taken to mean that the MP must supply both players with the same program: i.e., the only possible worlds are the ones in which both players have the same operating program. In this case, clearly, the MP would supply the players with a program that would cause them to both cooperate, because this would maximize the utility of Player 1 (whose interest the MP is supposed to be looking out for).

4.2 One-Shot Prisoner's Dilemma with Nonidentical Players

Next, with nonidentical players that have probability p of agreeing, we may calculate

$$U(C) = U(C \cap C, C)P(C, C|C) + U(C \cap C, D)P(C, D|C) = p \text{ payoff}(C, C) + (1-p)\text{payoff}(C, D)$$
$$U(D) = U(D \cap D, D)P(D, D|D) + U(D \cap D, C)P(D, C|D) = p \text{ payoff}(D, D) + (1-p)\text{payoff}(D, C)$$

So in this case, according to EDT, it is rational to cooperate if

$$U(C) > U(D)$$

which means (denoting e.g. $p_{CD} = \text{payoff}(C, D)$)

$$\begin{aligned} pp_{CC} + (1-p)p_{CD} &> pp_{DD} + (1-p)p_{DC} \\ p(p_{CC} - p_{DD}) &> (1-p)(p_{DC} - p_{CD}) \\ (p_{DC} - p_{CD}) &< r(p_{CC} - p_{DD}) \\ r &= p/(1-p) \end{aligned}$$

So for it to be rational to cooperate according to EDT,

$$X = \text{the added utility obtained from being the offender rather than the victim}$$

must be less than r times as big as

$$Y = \text{the added utility obtained from both cooperating rather than both defecting}$$

Note that if $p = .5$ then the two agents are independent, and in this case according to EDT one should cooperate only if $X < Y$

CDT, again, counsels that each player's choices have no impact on the other player during the course of the one-shot PD game, so that defecting is the right option.

What does CRDT say? Here the MP may be assumed to be operating under the constraint that, whatever program it supplies to Player 1, Player 2 will operate under *some* program that causes it to agree with Player 1 p of the time. The algebraic analysis given above applies perfectly well. Iff $X < \frac{p}{1-p}Y$ then MP should program Player 1 to cooperate. Thus, Player 1 should choose to cooperate.

5 When Is It Rational To Vote?

A similar analysis may be applied to the puzzle of deciding whether it's worth one's while to vote, via simply giving different interpretations to the variables in the above analysis of the PD. Consider a voter deciding whether to vote, in a situation where there is a group of voters who also have a decision whether or not to vote, and who would vote the same way as him (and each other) if they voted. Then we have, in the above algebra,

$$C = \text{vote}$$
$$D = \text{don't vote}$$

Say, for example, that $p = .9$, so that $p/(1-p) = 9$ (meaning, a "90% correlation" between the given voters, and the other voter or set of voters). Then, in order for voting to be a rational decision,

$X =$ The added utility obtained from not voting when others do, as opposed to voting when others do not ...

must be less than 9 times as big as ...

$$Y = \text{The added utility obtained from both voting versus neither voting}$$

Now, consider a voting situation involving one voter and set of M voters who are correlated with him, out of a total of N voters. Then, if $M << N$, clearly Y is very small, and unlikely to outweigh even a modest cost of voting (as incorporated in X). But, if M is close to N, then:

- X is very large (because if I don't vote and others do, the election will almost certainly still be won, since it's very unlikely I will be the tiebreaker)
- Y is also very large, but is about equal to X

So applying EDT, in this case it is rational to vote. And applying CRDT, one concludes that in this case the MP would reprogram the agent's brain with the propensity to vote, because according to the given constraints, this would result in a world in which everyone similar to the agent also had a propensity to vote.

For example, *if* one knows one is in a community of other CRDT fanatics, and believes they would take the same action as one would in a certain (sufficiently high) percentage of elections, then it may make sense to vote. Or, if one knows there is a group of people who

- tend to agree with one on the issues related to elections
- tend to vote as a group (for any reason, be it superrationality or just a herd mentality)

then it is also rational to vote. Of course, there is a tradeoff between the cost of voting (a cost in hassle, generally) versus the level of similarity between the voter and the group. As similarity goes to 1, the tolerable cost of voting becomes higher.

In order to apply this conclusion in practice, of course, one must determine how likely it is that the given constraint holds. The "voting may be useful" conclusion appears to hold up so long as one assumes the statistics regarding one's prior history are likely to govern one's current behavior.

6 Newcomb's Problem

This is perhaps the most vexing of all the decision-theoretic puzzles. The simplest version is as follows (lifted with edits from [Wik10a]):

A person is playing a game operated by an alien entity who is exceptionally skilled at predicting people's actions.

The player of the game is presented with two opaque boxes, labeled A and B. The player is permitted to take the contents of both boxes, or just of box B. Box A contains $1,000. The contents of box B, however, are determined as follows: At some point before the start of the game, the alien makes a prediction as to whether the player of the game will take just box B, or both boxes. If the alien predicts that both boxes will be taken, then box B will contain nothing. If the alien predicts that only box B will be taken, then box B will contain $1,000,000.

By the time the game begins, and the player is called upon to choose which boxes to take, the prediction has already been made, and the contents of box B have already been determined. That is, box B contains either $0 or $1,000,000 before the game begins, and once the game begins even the alien is powerless to change the contents of the boxes. Before the game begins, the player is aware of all the rules of the game, including the two possible contents of box B, the fact that its contents are based on the alien's prediction, and knowledge of the alien's infallibility. The only information withheld from the player is what prediction the alien made, and thus what the contents of box B are.

(In fact the original presentation of Newcomb's Problem was a little different than this – in it, the alien's predictions are only correct a certain percentage of the time; but this doesn't change the fundamental nature of the situation.)

EDT handles Newcomb's Problem straightforwardly, though controversially: one has

$$\begin{aligned}
U(\text{look in two}|\text{alien predicted look in two}) &= 1000 \\
U(\text{look in two }|\text{alien predicted look in one}) &= 1001000 \\
U(\text{look in one}|\text{alien predicted look in two}) &= 0 \\
U(\text{look in one}|\text{alien predicted look in one}) &= 1000000 \\
P(\text{alien predicted look in two }|\text{look in two}) &= 1 \\
P(\text{alien predicted look in one}|\text{look in one}) &= 1 \\
P(\text{alien predicted look in two}|\text{look in one}) &= 0 \\
P(\text{alien predicted look in one }|\text{look in two}) &= 0
\end{aligned}$$

and then

$$\begin{aligned}
U(\text{look in one}) &= 1000000 \\
U(\text{look in two}) &= 1000
\end{aligned}$$

There is not much subtlety here: if the agent really believes the alien's predictions are correct, and he is rational, then he will look in the closed box only.

This is somewhat similar to the PD situation with two clones; because exact cloning, like the predictive alien, essentially eliminates the hypothesis of free will right away, in the problem formulation.

As with PD, CDT is generally interpreted to give a different answer here. The basic idea is that

$$P(\text{alien predicted look in two} > \text{look in one}) = 0$$
$$P(\text{alien predicted look in one} > \text{look in one}) = 0$$
$$P(\text{alien predicted look in two} > \text{look in two}) = 0$$
$$P(\text{alien predicted look in one} > \text{look in two}) = 0$$

In other words, there is no causal relation between where one looks, and what the alien predicted, since the prediction occurred before the looking. So CDT recommends to look in both boxes.

What about CRDT? This is a trickier case than the previous ones!

One must, of course, assume the alien foresaw that the agent would be using CRDT to make his decision. So, the agent foresaw that the MP would be replacing the agent's brain with a program potentially having nothing to do with what came beforehand. Thus the game is between the MP and the alien, and is of the form:

- alien has predicted what MP will guess alien has predicted MP will guess alien has predicted...
- MP guesses what alien has predicted MP will guess alien has predicted MP will guess...

If we assume that the alien and the MP both have infinite computational capability, we get the equations

$$X = predict_A(Y)$$
$$Y = guess_{MP}(X)$$

where X and Y are both Boolean variables. If we assume that the alien and the MP are both correct in their predictions, then X and Y are unspecified by the above equation. Either $X = Y$ =two-boxes or $X = Y$ =one-box are valid solutions. These recursive equations, while correct, are not useful.

A slightly different situation occurs if one assumes the MP has finite computational capacity, whereas the alien is all-knowing. Then we have a finite situation of the form

- alien has predicted what MP will guess ... alien has predicted MP will guess alien has predicted
- MP guesses what alien has predicted MP will guess ... alien has predicted

and we may assume the alien's predictions are correct, whereas the MP's may or may not be. The MP knows it does not know what the alien predicts it will do. Here we have something very similar to the standard Newcomb's Problem, with the MP taking the place of the agent.

However, the difference from the standard Newcomb's Problem is that here the MP is assumed a rational deterministic agent, albeit with resource limitations (which may be assumed not too severe), which is single-mindedly dedicated to *maximizing the agent's expected reward across possible worlds*. The MP will estimate that, averaged over all possible worlds, the agent will get more reward from one-boxing than from two-boxing (because in each case the alien will correctly predict its doings). So the MP will choose to one-box.

What lets the MP escape the typical problems associated with Newcomb's Problem is that we have made very specific assumptions about its internal decision procedure. By assumption, the MP must work according to its algorithm, and the alien must predict that the MP will work according to its algorithm. So if the agent is following CRDT, it must obey the MP, which in this case will direct it to obey in an EDT-ish manner.

6.1 Nesov's Problem

An interesting variant of Newcomb's Problem was formulated by Vladimir Nesov [Nes10]:

Suppose the alien from Newcomb's Problem (who is known to be honest about how it poses these sorts of dilemmas) comes to you and says: "I just flipped a fair coin. I decided, before I flipped the coin, that if it

came up heads, I would ask you for $1000. And if it came up tails, I would give you $1,000,000 if and only if I predicted that you would give me $1000 if the coin had come up heads. The coin came up heads - can I have $1000?"

This is interesting, but doesn't pose any problems for the Master Programmer. The MP will assess that, averaged over possible worlds, the agent would do better to operate using a program that would pay the $1000.

7 The Psychopath Button

Finally, let us consider a problem due to [Ega07], which is known to be problematic for CDT.

Paul is debating whether to press the kill all psychopaths button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying. Should Paul press the button?

CDT tells Paul to press the button. This seems wrong! In this case EDT does better, and counsels Paul to keep himself alive and not press the button.

What does CRDT say? It says that, at the time referenced in the problem statement, the MP could maximize Paul's expected utility over possible worlds via supplying Paul with a control program that would *not* cause him to press the button. Thus, Paul should obey the MP's recommendation and not press the button.

Turning the decision over to the MP neatly nullifies Paul's inner debate over his potential psychopathy.

8 Conclusion

We have proposed Counterfactual Reprogramming Decision Theory, a novel alternative to evidential and causal decision theory, which appears to bypass the conceptual problems associated with these other approaches. CRDT constitutes a more direct confrontation of the problem of "choosing the action that will incur the greatest expected reward."

While we have shown CRDT to agree with human common sense in a variety of cases, we are not proposing it as a general model of human decision making. Humans tend not to act like rational utility maximizers according to *any* formal approach [GGK].

CRDT may be more plausible as a guide for AI decision making or automated decision support. In these contexts, however, one issue with CRDT is that implementing it in practice, in any direct and general way, would be computationally intractable. Here we have considered CRDT as a theoretical construct only, and shown it to handle various theoretical problem cases. Applying CRDT in real life contexts would involve the use of complex approximations to the posited Master Programmer, which leads into the domains of artificial general intelligence, cognitive architecture and automated program learning [GP05], and will not be enlarged upon further here.

A Toward a Formalization of CRDT

While this is only a semi-formal paper, in this brief Appendix we will outline a formalized version of CRDT, just to make clear what sort of theory this is (as it's a bit different on a mathematical-infrastructure level from EDT and CDT), and that all the above talk of Master Programmers and so forth isn't completely fanciful. The reader who lacks taste for mathematical formalism may skip this Appendix without significant loss; the above discussion does not depend on it in any particulars.

A.1 A Formal Model of Intelligent Agents

We will formalize CRDT within the formal agents framework used in [LH07b], in the context of formalizing the notion of general intelligence.

Consider a class of active agents which observe and explore their environment and also take actions in it, which may affect the environment. Formally, the agent sends information to the environment by sending symbols from some finite alphabet called the *action space* Σ; and the environment sends signals to the agent with symbols from an alphabet called the *perception space*, denoted $\mathcal{P}$. Agents can also experience rewards, which lie in the *reward space*, denoted $\mathcal{R}$, which for each agent is a subset of the rational unit interval.

The agent and environment are understood to take turns sending signals back and forth, yielding a history of actions, observations and rewards, which may be denoted

$$a_1 o_1 r_1 a_2 o_2 r_2 ...$$

or else

$$a_1 x_1 a_2 x_2 ...$$

if x is introduced as a single symbol to denote both an observation and a reward. The complete interaction history up to and including cycle t is denoted $ax_{1:t}$; and the history before cycle t is denoted $ax_{<t} = ax_{1:t-1}$.

The agent is represented as a function π which takes the current history as input, and produces an action as output. Agents need not be deterministic, an agent may for instance induce a probability distribution over the space of possible actions, conditioned on the current history. In this case we may characterize the agent by a probability distribution $\pi(a_t|ax_{<t})$. Similarly, the environment may be characterized by a probability distribution $\mu(x_k|ax_{<k}a_k)$. Taken together, the distributions π and μ define a probability measure over the space of interaction sequences.

It's interesting to specifically consider the class of environments that are *reward-summable*, meaning that the total amount of reward they return to any agent is bounded by 1. Where r_i denotes the reward experienced by the agent from the environment at time i, the *expected total reward* for the agent π from the environment μ is defined as

$$V_\mu^\pi \equiv E(\sum_1^\infty r_i) \leq 1$$

Generally speaking, more intelligent agents will achieve greater reward; but different agents may be better adapted to different environments.

A.2 Distributions over Possible Worlds

Next, introduce a second-order probability distribution ν, which is a probability distribution over the space of environments μ. The distribution ν assigns each environment a probability. One such distribution ν is the Solomonoff-Levin universal prior distribution in which one sets $\nu = 2^{-K(\mu)}$; but this is not the only distribution ν of interest.

Denote an alteration of an interaction sequence $I_{s,t}^a$ for an agent a by $\tilde{I}_{s,t}^a$, and the set of all such altered interaction sequences for agent a by $\mathcal{I}^a$. It is interesting to look at those cases for which $d_I(I_{s,t}^a, \tilde{I}_{s,t}^a)$ is small, where $d_I(\cdot,\cdot)$ is some (perhaps rescaled version of) some measure of sequence similarity, such as neighborhood correlation or PSI-BLAST. Using this one may construct a probability distribution ν over environments μ that tends to give larger probabilities to nearby sequences, as measured by the chosen similarity measure. This is one way of constructing a set of probabilistically weighted "possible worlds."

These two approaches (prior based and similarity based) may of course be combined, as well. Given a prior distribution ν_1 over environments, and another distribution ν_2 defined based on $d_I(I_{g,s,t}^a, \tilde{I}_{s,t}^a)$, one may look at the conjunction $\nu = \nu_1 \cap \nu_2$. Conceptually, this represents a distribution that values possible worlds close to the given interaction sequence, but restricts attention to those possible worlds that are considered viable according to the prior distribution.

A.3 A Partially Reprogrammable Agent

So far no commitment has been made about the internal structure of the intelligent agent. However, CRDT deals with agents that are in some sense *partially reprogrammable*. For instance, we may assume that the agent has associated with it some finite amount of memory, and some finite amount of processing time per

cycle (where a cycle consists of a single perception, action and reward), that is reprogrammable, i.e. that may be filled with arbitrary "code" or computational content.

Qualitatively, when we say the agent is "choosing" something, we may understand that this choice is related to changes in the code in the agent's reprogrammable sector. And in CRDT, the Master Programmer is understood to modify the code in this sector.

A.4 The Master Programmer

The MP in CRDT, then, may be defined as a system that, at any given time,

- has access to the complete history of the agent a, including the agent's internal state
- has the goal of finding code to put into the agent's reprogrammable sector, that will maximize a's expected reward, as averaged over possible environments according to some distribution ν (e.g. constructed as outlined above)
- has a very large amount of processing power and memory at its disposal (one may also consider the case where the MP has infinite processing power and memory; but this is unnecessary for the points to be made in this paper)
- pursues this goal according to the $AIXI^{tl}$ algorithm [Hut05] (or something similar)

If *all* of the agent were reprogrammable, and the MP were actually allowed to reprogram the agent, and the agent had a very large amount of processing power and memory – then as both the agent's and MP's processing power and memory tended to infinity, the agent would tend toward optimal intelligence as defined in [LH07b].

But even if the agent doesn't want to let the MP reprogram itself, or doesn't have access to any MP, it can still use the notion of an MP as a guide for its actions – asking what it would do in a given situation, if it were to allow an MP to replace its reprogrammable sector with whatever code it wanted, based on its goal of maximizing a's expected reward. This is the basic idea of CRDT.

References

[Dea09] Wei Dai and et al. Toward a new decision theory. *Less Wrong*, 2009. `http://lesswrong.com/lw/15m/my_timeless_decision_theory/`.

[Ega07] Andy Egan. Some counterexamples to causal decision theory. *Philosophical Review*, 116, 2007.

[GGK] Thomas Gilovich, Dale Griffin, and Dale Kahneman. *Heuristics and Biases*. Cambridge University Press.

[GH81] A. Gibbard and W.L Harper. Counterfactuals and two kinds of expected utility. pages 153–190, 1981.

[GP05] Ben Goertzel and Cassio Pennachin. *Artificial General Intelligence*. Springer, 2005.

[Hof79] Douglas Hofstadter. *Godel, Escher, Bach: An Eternal Golden Braid*. Basic, 1979.

[Hut05] Marcus Hutter. *Universal AI*. Springer, 2005.

[Lew81] D. Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59:5–30, 1981.

[LH07a] Shane Legg and Marcus Hutter. A collection of definitions of intelligence. pages –. IOS, 2007.

[LH07b] Shane Legg and Marcus Hutter. A definition of machine intelligence. *Minds and Machines*, 17:–, 2007.

[Nes10] Counterfactual mugging. *Less Wrong*, 2010. `http://lesswrong.com/lw/3l/counterfactual_mugging/`.

[Wik10a] Newcomb's problem. *Wikipedia*, 2010.

[Wik10b] Prisoner's dilemma. *Wikipedia*, 2010.