Law as the most comprehensible guide for morality.
To Address Corrigibility and Value Learning

(Please don't dismiss the following message because of the way I express myself. English is my second language)

Background:
I have a double degree in common law (used by United States and the Commonwealth) and civil law (used by the rest of the world) and and I have studied the development and history of law, which gives me good insights in the field. I gave at talk at a conference on how law and morality are tightly connected. Usually, the law is representing the morality of a country with few years of delay.

The foundation of law is actually a set of rules of priority to know how to interpret all the laws and decisions. So in theory, when the deep learning system will be more advanced, it could be the code for deep learning into a database. There is one online at www.commonlii.org/ for the commonwealth, https://www.law.cornell.edu/ for the USA or a private system if we don't want the AI to be connected to internet such as LexisNexis.

Why I believe coding the hierarchy of law would be a great way for AI:

1. Because this is the biggest pool of writing to describe in detail why we take ta moral decision over another. Judges are there to put in place a system that is the fairest possible while maintaining order/avoiding chaos. They are an elite who are supposed to be moral. So the system of law is based on thousands of men who build on others knowledge and experience. It is unprecedented anywhere else, not even all the writing of the philosophers in ethics. (Maybe we could already with a Deep Learning program classify the values which are important to us based on the recurrence of them in the judgements).
2. It would allow to adjust anything going wrong with the AI. For example, the AI starts doing paper clips abnormally. The supreme court could do an express decision or the legislation could do an express law stating that no more paper clips can be produced until this law is repealed or this decision is reversed. The goal of the AI would be after the one of respecting the Law, thus AI would be a better citizen than us. Any loophole that the AI would find could be addressed by the legislation as a new law.
3. It would allow that the AI would always follow the humans morality.

In common law, the hierarchy is

a) Constitution,
The constitution is interpreted by the courts, so is guided by the precedents, but if there is a conflict between the law and the constitution, the constitution will trump laws.
b) Laws

The laws are also interpreted by the courts, so guided by the precedents.
c) Precedents.
The precedents have also a hierarchy.

Judges have to follow the precedents by

1. The Supreme court,
2. if there is no precedent on the subject, by the court of appeal from the same state.

BUT the three following principles apply:

i) The newest decision is higher on the hierarchy,
ii) the decisions that has not been reversed are also higher (More a case is reversed, less weight it has),
iii) More a case is followed, or cited, more the case has weight.

The last 3 principles also apply to the following. The following decisions can influence the judges but are not binding (also in order of hierarchy):

3. Decisions by the lower courts (name changes between states and countries) from the same state
4. By the court of Appeal from the main state from that country
5. By the court of Appeal from the states the most similar to that state in the country (4 and 5 can be reversed, depending of the Ministry of Justice policy)
6. By the court of Appeal from any other states in the country
7. By the lower courts from the main state from that country
8. By the lower courts from the states the most similar to that state in the country (7 and 8 can be reversed, depending of the Ministry of Justice policy)
9. By the lower courts of from any other states in the country
10. By any supreme courts from other countries.

The main problem with the common law is that it is very intuitive. Decisions are hundreds of pages and people have to find the 'ratio decidendi' which means the rule of that decision. Usually, it's the part that is repeated by other decisions. However, any other parts can be used to justify something. As well, there is the descent opinion which may have some weight to reverse the decision in the future.

Another problem is that judges have more power. In the United States, where the judges are bi-partisan, the future is more uncertain (for example if Trump is elected and nominate extreme republican). Canada, Australia or England have more moral laws, more moderate.

On civil law, on the other hand, things are 'simpler'.

1. Constitution
2. Law (much more detailed, it's like if all the 'ratio decidendi' were coded (civil code).
3. Jurisprudence. (No decisions are binding, it is only influencing judges (same hierarchy than in common law and same principles)

There are advantages or disadvantages to both (I could talk to you longer about it if you are interested).
This way, if the AI respects all the laws, they would be at least as moral and ‘friendly’ as the average of all the judges of the supreme court.