Addleton Academic Publishers

Analysis and Metaphysics

GENERAL PURPOSE INTELLIGENCE: ARGUING THE ORTHOGONALITY THESIS

STUART ARMSTRONG
stuart.armstrong@philosophy.ox.ac.uk
Future of Humanity Institute, Oxford Martin School
Philosophy Department, University of Oxford

In his paper "The Superintelligent Will", Nick Bostrom formalised the Orthogonality thesis: the idea that the final goals and intelligence levels of artificial agents are independent of each other. This paper presents arguments for a (narrower) version of the thesis. It proceeds through three steps. First it shows that superintelligent agents with essentially arbitrary goals can exist in our universe – both as theoretical impractical agents such as AIXI and as physically possible real-world agents. Then it argues that if humans are capable of building human-level artificial intelligences, we can build them with an extremely broad spectrum of goals. Finally it shows that the same result holds for any superintelligent agent we could directly or indirectly build. This result is relevant for arguments about the potential motivations of future agents: knowing an artificial agent is of high intelligence does not allow us to presume that it will be moral, we will need to figure out its goals directly.

Keywords: AI; Artificial Intelligence; efficiency; intelligence; goals; orthogonality

1 The Orthogonality thesis

Scientists and mathematicians are the stereotypical examples of high intelligence humans. But their morality and ethics have been all over the map. On modern political scales, they can be left- (Oppenheimer) or right-wing (von Neumann) and historically they have slotted into most of the political groupings of their period (Galois, Lavoisier). Ethically, they have ranged from very humanitarian (Darwin, Einstein outside of his private

life), through amoral (von Braun) to commercially belligerent (Edison) and vindictive (Newton). Few scientists have been put in a position where they could demonstrate genuinely evil behaviour, but there have been a few of those (Teichmüller, Philipp Lenard, Ted Kaczynski, Shirō Ishii).

Of course, many scientists have been absolutely conventional in their views and attitudes given the society of their time. But the above examples hint that their ethics are not strongly impacted by their high intelligence; intelligence and ethics seem 'orthogonal' (varying independently of each other, to some extent). If we turn to the case of (potential) artificial intelligences we can ask whether that relation continues: would high intelligence go along with certain motivations and goals, or are they unrelated?

To avoid the implicit anthropomorphisation in terms such as 'ethics', we will be looking at agents 'final goals' – the ultimate objectives they are aiming for. Then the Orthogonality thesis, due to Nick Bostrom (Bostrom, 2012), states that:

> *Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.*

It is analogous to Hume's thesis about the independence of reason and morality (Hume, 1739), but applied more narrowly, using the normatively thinner concepts 'intelligence' and 'final goals' rather than 'reason' and 'morality'.

But even 'intelligence', as generally used, has too many connotations. A better term would be efficiency, or instrumental rationality, or the ability to effectively solve problems given limited knowledge and resources (Wang, 2011). Nevertheless, we will be sticking with terminology such as 'intelligent agent', 'artificial intelligence' or 'superintelligence', as they are well established, but using them synonymously with 'efficient agent', artificial efficiency' and 'superefficient algorithm'. The relevant criteria is whether the agent can effectively achieve its goals in general situations, not whether its inner process matches up with a particular definition of what intelligence is.

Thus an artificial intelligence (AI) is an artificial algorithm, deterministic or probabilistic, implemented on some device, that demonstrates an ability to achieve goals in varied and general situations[1]. We don't assume that it need be a computer program, or a well laid-out algorithm with clear loops and structures – artificial neural networks or evolved genetic algorithms certainly qualify.

A human level AI is defined to be an AI that can successfully accomplish any task at least as well as an average human would (to avoid worrying about robot bodies and such-like, we may restrict the list of tasks to those accomplishable over the internet). Thus we would expect the AI to hold conversations about Paris Hilton's sex life, to compose ironic limericks, to shop for the best deal on Halloween costumes and to debate the proper role of religion in politics, at least as well as an average human would.

A superhuman AI is similarly defined as an AI that would exceed the ability of the best human in all (or almost all) tasks. It would do the best research, write the most successful novels, run companies and motivate employees better than anyone else. In areas where there may not be clear scales (what's the world's best artwork?) we would expect a majority of the human population to agree the AI's work is among the very best.

Nick Bostrom's paper argued that the Orthogonality thesis does not depend on the Humean theory of motivation, but could still be true under other philosophical theories. It should be immediately apparent that the Orthogonality thesis is related to arguments about moral realism. Despite this, we will not address the fertile and extensive literature on this subject. Firstly, because it is contentious: different schools of philosophical thought have different interpretations of the truth and meaning of moral realism, disputes that cannot be currently resolved empirically. Since we are looking to resolve a mainly empirical question – what systems of motivations could

[1] We need to assume it has goals, of course. Determining whether something qualifies as a goal-based agent is very tricky (researcher Owain Evans is trying to establish a rigorous definition), but this paper will adopt the somewhat informal definition that an agent has goals if it achieves similar outcomes from very different starting positions. If the agent ends up making ice cream in any circumstances, we can assume ice creams are in its goals.

we actually code into a putative AI – this theoretical disagreement is highly problematic.

Secondly, we hope that by approaching the issue from the computational perspective, we can help shed new light on these issues. After all, we do not expect that the trigger mechanism of a cruise missile to block detonation simply because people will die – but would an "ultra-smart bomb" behave the same way? By exploring the goals of artificial systems up to higher level of efficiency, we may contribute to seeing which kinds of agents are susceptible to moral realism arguments, and which are not.

Thus this paper will content itself with presenting direct arguments for the Orthogonality thesis. We will assume throughout that human level AIs (or at least human comparable AIs) are possible (if not, the thesis is void of useful content). We will also take the position that humans themselves can be viewed as non-deterministic algorithms[2]: this is not vital to the paper, but is useful for comparison of goals between various types of agents. We will do the same with entities such as committees of humans, institutions or corporations, if these can be considered to be acting in an agent-like way.

The thesis itself might be critiqued for over-obviousness or triviality – a moral anti-realist, for instance, could find it too evident to need defending. Nevertheless, the argument that AIs – or indeed, any sufficiently intelligent being – would necessarily behave morally is a surprisingly common one. A. Kornai, for instance, considers it as a worthwhile starting point for investigations into AI morality (Kornai, 2013). He bases his argument on A. Gewith's approach in his book, Reason and Morality (Gewirth, 1978) (the book's argument can be found in a summarised form in one of E. M. Adams's papers (Adams, 1980)) in which it is argued that all agents must follow a "Principle of Generic Consistency" that causes them to behave in accordance with all other agent's generic rights to freedom and well-being. Others have argued that certain specific moralities are attractors in the space of moral systems, towards which any AI will tend if they start off with certain mild constraints (Waser, 2008). Because of these and other examples

[2] Since every law of nature is algorithmic (with some probabilistic process of known odds), and no exceptions to these laws are known, neither for human nor non-human processes.

(and some online criticism of the Orthogonality thesis[3]), we thought the thesis was worth defending explicitly, and that the argument brought out in its favour would be of general interest to the general discussion.

1.1 Qualifying the Orthogonality thesis

The Orthogonality thesis, taken literally, is false. Some motivations are mathematically incompatible with changes in intelligence ("I want to prove the Gödel statement for the being I would be if I were more intelligent"). Some goals specifically refer to the intelligence of the agent, directly ("I want to be much less efficient!") or indirectly ("I want to impress people who want me to be much less efficient!"). Though we could make a case that an agent wanting to be less efficient could initially be of any intelligence level, it won't stay there long, and it's hard to see how an agent with that goal could have become intelligent in the first place. So we will exclude from consideration those goals that intrinsically refer to the intelligence level of the agent.

We will also exclude goals that are so complex or hard to describe that the complexity of the goal becomes crippling for the agent. If the agent's goal takes five planets worth of material to describe, or if it takes the agent twenty years each time it checks what its goal is, then it's obvious that that agent can't function as an intelligent being on any reasonable scale.

Many have made the point that there is likely to be convergence in *instrumental* goals (Omohundro, 2008). Whatever their final goals, it would generally be in any agent's interest to accumulate more power, to become more intelligence and to be able to cooperate with other agents of similar ability (and to have all the negotiation, threatening and cajoling skills that go along with that cooperation). Note the similarity with what John Rawls called 'primary goods' (Rawls, 1971). We will however be focusing

[3] See for example http://philosophicaldisquisitions.blogspot.co.uk/2012/04/bostrom-on-superintelligence-and.html, which criticise the thesis specifically.

exclusively on final goals, as the instrumental goals are merely tools to accomplish these[4].

Further we will not try to show that intelligence and final goals can *vary* freely, in any dynamical sense (it could be quite hard to define this). Instead we will look at the thesis as talking about possible states: that there exist agents of all levels of intelligence with any given goals. Since it's always possible to make an agent stupider or less efficient, what we are really claiming is that there could exist possible high-intelligence agents with any given goal. Thus the restricted Orthogonality thesis that we will be discussing is:

High-intelligence agents can exist having more or less any final goals (as long as these goals are of feasible complexity, and do not refer intrinsically to the agent's intelligence)[5].

We will be looking at two variations of the "can exist" clause: whether the agent can exist in theory, and whether we could build such an agent (given that we could build an AI at all). Though evidence will be presented directly for this thesis in the theoretic agent case, the results of this paper cannot be considered to "prove" the thesis for agents we could build (though they certainly raise its likelihood). In that case, we will be looking at proving a still weaker thesis:

The fact of being of high intelligence provides extremely little constraint on what final goals an agent could have (as long as these goals are of feasible complexity, and do not refer intrinsically to the agent's intelligence).

[4] An AI skilled in cooperation would drop this if cooperation no longer served its purpose; similarly, an AI accumulating power and resources would stop doing this if it found better ways of achieving its goals.

[5] Even logically impossible goals can exist: "construct a disproof of Modus Ponens (within classical logic)" is a perfectly fine goal for an intelligence to have – it will quickly realise that this translates to "prove classical logic is inconsistent", a task mathematicians have occasionally attempted.

That thesis still has nearly all the relevant practical implications that the strong Orthogonality thesis does.

1.2 Orthogonality in practice for AI designers

The arguments presented in this paper are all theoretical. They posit that AIs with certain goals either 'can exist', or that 'if we could build an AI, we could build one with any goal'. In practice, the first AIs, if and when they are created, will be assembled by a specific team, using specific methods, and with specific goals in mind. They may be more or less successful at inculcating the goals into the AI (or, as is common in computer programming, they may inculcate the goals exactly, only to realise later that these weren't the goals they really wanted). The AI may be trained by interacting with certain humans in certain situations, or by understanding certain ethical principles, or by a myriad of other possible methods, which will likely focus on a narrow target in the space of goals. The relevance of the Orthogonality thesis for AI designers is therefore mainly limited to a warning: that high intelligence and efficiency are not enough to guarantee positive goals, and that they thus need to work carefully to inculcate the goals they value into the AI.

2 Orthogonality for theoretic agents

If we were to step back for a moment and consider, in our mind's eyes, the space of every possible algorithm, peering into their goal systems and teasing out some measure of their relative intelligences, would we expect the Orthogonality thesis to hold? Since we are not worrying about practicality or constructability, all that we would require is that for any given goal system (within the few constraints enumerated above), there exists a theoretically implementable algorithm of high intelligence.

Any measurable[6] goal can be paired up with a reward signal: an agent gets a reward for achieving states of the world desired by the goal, and denied these rewards for actions that fail to do so. Among reward signal

[6] Measuring a goal brings up subtle issues with the symbol grounding problem and similar problems. We'll ignore these issues in the present paper.

maximisers, the AIXI is the theoretically best agent there is, more successful at reaching its goals (up to a finite constant) than any other agent (Hutter, 2005). AIXI itself is incomputable, but there are computable variants such as AIXItl or Gödel machines (Schmidhuber, 2007) that approximate AIXI's efficiency. These methods work for whatever reward signal plugged into them. Or we could simply imagine a supercomputer with arbitrarily large amounts of computing power and a decent understanding of the laws of physics (a 'Laplace demon' (Laplace, 1814) capable of probabilistic reasoning), placed 'outside the universe' and computing the future course of events. Paired with an obedient active agent inside the universe with a measurable goal, for which it would act as an advisor, this would also constitute an 'ultimate agent'. Thus in the extreme theoretical case, the Orthogonality thesis seems true.

There is only one problem with these agents: they are either impossible in practice (AIXI or Laplace's demon), or require incredibly large amounts of computing resources to work. Let us step down from the theoretical pinnacle and require that these agents could actually exist in our world (still not requiring that we be able or likely to build them).

An interesting thought experiment occurs here. We could imagine an AIXI-like super-agent, with all its impractical resources, that is tasked to design and train an AI that could exist in our world, and that would accomplish the super-agent's goals. Using its own vast intelligence, the super-agent would therefore design a constrained agent maximally effective at accomplishing those goals in our world. Then this agent would be the high-intelligence real-world agent we are looking for. It doesn't matter than the designer is impossible in practice – if the super-agent can succeed in the theoretical thought experiment, then the trained AI can exist in our world.

This argument generalises to other ways of producing the AI. Thus to deny the Orthogonality thesis is to assert that there is a goal system G, such that, among other things:

1. There cannot exist any efficient real-world algorithm with goal G.
2. If a being with arbitrarily high resources, intelligence, time *and* goal G, were to try design an efficient real-world algorithm with the same goal, it must fail.

3. If a human society were highly motivated[7] to design an efficient real-world algorithm with goal G, and were given a million years to do so along with huge amounts of resources, training and knowledge about AI, it must fail.

4. If a high-resource human society were highly motivated to achieve the goal G, then it could not do so (here the human society itself is seen as the algorithm).

5. Same as above, for any hypothetical alien societies.

6. There cannot exist *any* pattern of reinforcement learning that would train a highly efficient real-world intelligence to follow the goal G.

7. There cannot exist *any* evolutionary or environmental pressures that would evolve a highly efficient real world intelligences following goal G.

All of these seem extraordinarily strong claims to make! The last claims all derive from the first, and merely serve to illustrate how strong the first claim actually is. Claim 4, in particular, seems to run counter to everything we know about human nature.

3 Orthogonality for human-level AIs

Of course, even if efficient agents could exist for all these goals, that doesn't mean that we could ever build them, even if we could build AIs. In this section, we'll look at the ground for assuming the Orthogonality thesis holds for human-level agents. Since intelligence isn't varying much, the thesis becomes simply:

If we could construct human-level AIs at all, then there is extremely little constraint on the final goals that such AIs could have (as long as these goals are of feasible complexity, and do not refer intrinsically to the agent's intelligence).

So, is this true? The arguments in this section are generally independent of each other, and can be summarised as:

[7] A motivation might simply be a threat: some truthful powerful being saying "Design an algorithm with goal G. If you succeed, I will give you great goods; if you fail, I will destroy you all. The algorithm will never be used in practice, so there are no moral objections to it being designed."

1. Some possible AI designs have orthogonality built right into them.
2. AI goals can reach the span of human goals, which is large.
3. Algorithms can be combined to generate an AI with any easily measurable goal.
4. Various algorithmic modifications can be used to further expand the space of possible goals, if needed.

3.1 Utility functions

One classical picture of a rational agent is of an agent with a specific utility function, which it will then act to maximise in expectation. This picture encapsulates the Orthogonality thesis: whatever the utility function, the rational agent will then attempt to maximise it, using the approaches in all cases (planning, analysing input data, computing expected results). If an AI is built according to this model, with the utility function being prescriptive (given to the AI in a program) rather than descriptive (an abstract formalisation of an agent's other preferences), then the thesis would be trivially true: we could simply substitute the utility function for whichever one we desired.

However, many putative agent designs are not utility function based, such as neural networks, genetic algorithms, or humans. So from now on we will consider that our agents are not expected utility maximisers with clear and separate utility functions, and look at proving Orthogonality in these harder circumstances.

3.2 The span of human motivations

It seems a reasonable assumption that if there exists a human being with particular goals, and we can program an AI, then we can construct a human-level AI with similar goals. This is immediately the case if the AI was a whole brain emulation/upload (Sandberg & Bostrom, 2008), a digital copy of a specific human mind. Even for more general agents, such as evolved agents, this remains a reasonable thesis. For a start, we know that real-world evolution has produced us, so constructing human-like agents that way is certainly possible. Human minds remain our only real model of general intelligence, and this strongly directs and informs our AI designs, which are likely to be as human-similar as we can make them. Similarly, human goals

are the easiest goals for us to understand, hence the easiest to try and implement in AI. Hence it seems likely that we could implement most human goals in the first generation of human-level AIs.

So how wide is the space of human motivations[8]? Our race spans foot-fetishists, religious saints, serial killers, instinctive accountants, role-players, self-cannibals, firefighters and conceptual artists. The autistic, those with exceptional social skills, the obsessive compulsive and some with split-brains. Beings of great empathy and the many who used to enjoy torture and executions as public spectacles[9]. It is evident that the space of possible human motivations is vast[10]. For any desire, any particular goal, no matter how niche[11], pathological, bizarre or extreme, as long as there is a single human who ever had it, we could build and run an AI with the same goal.

But with AIs we can go even further. We could take any of these goals as a starting point, make them malleable (as goals are in humans), and push them further out. We could provide the AIs with specific reinforcements to push their goals in extreme directions (reward the saint for ever-more saintly behaviour). If the agents are fast enough, we could run whole societies of them with huge varieties of evolutionary or social pressures, to further explore the goal-space.

---

[8] One could argue that we should consider the space of general animal intelligences – octopuses, supercolonies of social insects, etc... But the methods described can already produce these animal's types of behaviours.

[9] Even today, many people have had great fun torturing and abusing their characters in games like "the Sims" (http://meodia.com/article/281/sadistic-ways-people-torture-their-sims/). The same urges are present, albeit diverted to fictionalised settings. Indeed games offer a wide variety of different goals that could conceivably be imported into an AI if it were possible to erase the reality/fiction distinction in its motivation.

[10] As can be shown by a glance through a biography of famous people – and famous means they were generally allowed to rise to prominence in their own society, so the space of possible motivations was already cut down.

[11] Of course, if we built an AI with that goal and copied it millions of times, it would no longer be niche.

We may also be able to do surgery directly on their goals, to introduce more yet variety. For example, we could take a dedicated utilitarian charity worker obsessed with saving lives in poorer countries (but who doesn't interact, or want to interact, directly with those saved), and replace 'saving lives' with 'maximising the number of paperclips in the universe' or any similar abstract goal. This is more speculative, of course – but there are other ways of getting similar results.

3.3 Instrumental goals as final goals

If someone were to hold a gun to your head, they could make you do almost anything. Certainly there are people who, with a gun at their head, would be willing to do almost anything. A distinction is generally made between instrumental goals and final goals, with the former being seen as simply paths to the latter, and interchangeable with other plausible paths. The gun to your head disrupts the balance: your final goal is simply not to get shot, while your instrumental goals become what the gun holder wants them to be, and you put a great amount of effort into accomplishing the minute details of these instrumental goals. Note that the gun has not changed your level of intelligence or ability.

This is relevant because instrumental goals seem to be far more varied in humans than final goals. One can have instrumental goals of filling papers, solving equations, walking dogs, making money, pushing buttons in various sequences, opening doors, enhancing shareholder value, assembling cars, bombing villages or putting sharks into tanks. Or simply doing whatever the guy with gun at our head orders us to do. If we could accept human instrumental goals as AI final goals, we would extend the space of goals quite dramatically.

To do so we would want to put the threatened agent, and the gun wielder, together into the same AI. Algorithmically there is nothing extraordinary about this: certain subroutines have certain behaviours depending on the outputs of other subroutines. The 'gun wielder' need not be particularly intelligent: it simply needs to be able to establish whether its goals are being met. If for instance those goals are given by a utility function then all that is required in an automated system that measure progress toward increasing utility and punishes (or erases) the rest of the AI if not. The 'rest of AI' is just required to be a human-level AI which would be susceptible to this kind of pressure. Note that we do not require that it

even be close to human in any way, simply that it place a highest value on self-preservation (or on some similar small goal that the 'gun wielder' would have power over).

For humans, another similar model is that of a job in a corporation or bureaucracy: in order to achieve the money required for their final goals, some human are willing to perform extreme tasks (organising the logistics of genocides, weapon design, writing long emotional press releases they don't agree with at all). Again, if the corporation-employee relationship can be captured in a single algorithm, this would generate an intelligent AI whose goal is anything measurable by the 'corporation'. The 'money' could simply be an internal reward channel, perfectly aligning the incentives.

If the subagent is anything like a human, they would quickly integrate the other goals into their own motivation[12], removing the need for the gun wielder/corporation part of the algorithm.

3.4 Noise, anti-agents and goal combination

There are further ways of extending the space of goals we could implement in human-level AIs. One simple way is simply to introduce noise: flip a few bits and subroutines, add bugs and get a new agent. Of course, this is likely to cause the agent's intelligence to decrease somewhat, but we have generated new goals. Then, if appropriate, we could use evolution or other improvements to raise the agent's intelligence again; this will likely undo some, but not all of effect of the noise. Or we could use some of the tricks above to make a smarter agent implement the goals of the noise-modified agent.

A more extreme example would be to create an anti-agent: an agent whose single goal is to stymie the plans and goals of single given agent. This already happens with vengeful humans, and we would just need to dial it up: have an anti-agent that would do all it can to counter the goals of a given agent, even if that agent doesn't exist ("I don't care that you're dead, I'm still going to despoil your country, because that's what you'd wanted me to not do"). This further extends the space of possible goals.

[12] Such as the hostages suffering from Stockholm syndrome (de Fabrique, Romano, Vecchi, & van Hasselt, 2007).

Different agents with different goals can also be combined into a single algorithm. With some algorithmic method for the AIs to negotiate their combined objective and balance the relative importance of their goals, this procedure would construct a single AI with a combined goal system. There would likely be no drop in intelligence/efficiency: committees of two can work very well towards their common goals, especially if there is some automatic penalty for disagreements.

3.5 Further tricks up the sleeve

This section started by emphasising the wide space of human goals, and then introduced tricks to push goal systems further beyond these boundaries. The list isn't exhaustive: there are surely more devices and ideas one can use to continue to extend the space of possible goals for human-level AIs. Though this might not be enough to get every goal, we can nearly certainly use these procedures to construct a human-level AI with any human-comprehensible goal. But would the same be true for superhuman AIs?

4 Orthogonality for superhuman AIs

We now come to the area where the Orthogonality thesis seems the most vulnerable. It is one thing to have human-level AIs, or abstract superintelligent algorithms created ex nihilo, with certain goals. But if ever the human race were to design a superintelligent AI, there would be some sort of process involved – directed evolution, recursive self-improvement[13], design by a committee of AIs, or similar – and it seems at least possible that such a process could fail to fully explore the goal-space. The Orthogonality thesis in this context is:

If we could construct superintelligent AIs at all, then there is extremely little constraint on the final goals that such AIs could have (as long as these goals are of feasible complexity, and do not refer intrinsically to the agent's intelligence).

There are two counter-theses. The weakest claim is:

[13] See for instance E. Yudkowsky's design "General Intelligence and Seed AI 2.3" http://singinst.org/ourresearch/publications/GISAI/

Incompleteness: there are large categories of goals that no superintelligence designed by us could have.

A stronger claim is:

Convergence: all human-designed superintelligences would have one of a small set of goals.

Here 'small' means 'smaller than the space of current human motivations', thus very small in comparison with the space of possible AI goals. They should be distinguished; Incompleteness is all that is needed to contradict Orthogonality, but Convergence is often the issue being discussed. Often Convergence is stated in terms of a particular model of metaethics, to which it is assumed all agents will converge (see some of the references in the introduction, or various online texts and argument[14]).

4.1 No convergence

The plausibility of the convergence thesis is highly connected with the connotations of the terms used in it. "All human-designed rational beings would follow the same morality (or one of small sets of moralities)" sounds plausible; in contrast "all human-designed superefficient algorithms would accomplish the same task" seems ridiculous. To quote an online commentator, how good at playing chess would a chess computer have to be before it started feeding the hungry?

Similarly, if there were such a convergence, then all self-improving or constructed superintelligence must fall prey to it, even if it were actively seeking to avoid it. After all, the self-improving lower-level AIs or the designers have certain goals in mind (as we've seen in the previous section, if the designers are AIs themselves, they could have potentially any goals in mind). Obviously, they would be less likely to achieve their goals if these goals were to change as they got more intelligent (Omohundro, 2008) (see also N. Bostrom's forthcoming book *Superintelligence: Groundwork to a Strategic Analysis of the Machine Intelligence Revolution*). The same goes if the superintelligent AI they designed didn't share these goals. Hence the AI designers will be actively trying to prevent such a convergence, if they suspected that one was likely to happen. If for instance their goals were

[14] Such as J. Müller's "Ethics, risks and opportunities of superintelligences" http://www.jonatasmuller.com/superintelligences.pdf

immoral, they would program their AI not to care about morality; they would use every trick up their sleeves to prevent the AI's goals from drifting from their own.

So the convergence thesis requires that for the vast majority of goals G:

1. It is possible for a superintelligence to exist with goal G (by section 2).
2. There exists an entity with goal G (by section 3), capable of building a superintelligent AI.
3. Yet any attempt of that entity to build a superintelligent AI with goal G will be a failure, and the superintelligence's goals will converge on some other goal.
4. This is true even if the entity is aware of the convergence and explicitly attempts to avoid it.
5. If the superintelligence were to be constructed by successive self-improvement, then an entity with goal G operating on itself to boost its intelligence is unable to do so in a way that would preserve goal G.

This makes the convergence thesis very unlikely. The argument also works against the incompleteness thesis, but in a weaker fashion: it seems more plausible that some types of goals would be unreachable, despite being theoretically possible.

There is another interesting aspect of the convergence thesis: these goals G are to emerge, somehow, without them being aimed for or desired. If one accepts that goals aimed for will not be reached, one has to ask why convergence is assumed: why not divergence? Why not assume that though G is aimed for, random accidents or faulty implementation will lead to the AI ending up with one of a much wider array of possible goals, rather than a much narrower one? We won't delve deeper into this, and simply make the point that "superintelligent AIs won't have the goals we want them to have" is therefore not an argument in favour of the convergence thesis.

4.2 Oracles show the way

If the Orthogonality thesis is wrong, then it implies that Oracles are impossible to build. An Oracle is a superintelligent AI that accurately answers human questions about the world, such as the likely consequences of certain policies and decisions (Armstrong, Sandberg, & Bostrom,

2012)[15]. If such an Oracle could be built, then we could attach it to a human-level AI with goal G. The human-level AI could then ask the Oracle what the results of different decisions actions could be, and choose the action that best accomplishes G. In this way, the combined system would be a superintelligent AI with goal G.

What makes the "no Oracle" implication even more counterintuitive is that *any* superintelligence must be able to look ahead, design actions, predict the consequences of its actions, and choose the best one available. But the convergence and indifference theses imply that this general skill is one that we can make available *only* to AIs with certain specific goals. Though agents with those specific goals are capable of doing effective predictions, they automatically lose this ability if their goals were to change.

4.3 Tricking the controller

Just as with human-level AIs, one could construct a superintelligent AI by wedding together a superintelligence with a large motivated committee of human-level AIs dedicated to implementing a goal G, and checking the superintelligence's actions. Thus to deny the Orthogonality thesis requires that one believes that the superintelligence is always capable of tricking this committee, no matter how detailed and thorough their oversight.

This argument extends the Orthogonality thesis to moderately superintelligent AIs, or to any situation where there's a diminishing return to intelligence. It only fails if we take AI to be fantastically superhuman: capable of tricking or seducing any collection of human-level beings.

4.4 Temporary fragments of algorithms, fictional worlds and extra tricks

These are other tricks that can be used to create an AI with any goals. For any superintelligent AI, there are certain inputs that will make it behave in certain ways. For instance, a human-loving moral AI could be compelled

[15] Not to be confused with the concept of Oracle in computer science, which is either an abstract machine capable of instantaneous computations in various complexity classes, or mechanism in software testing.

to follow most goals G for a day, if they were rewarded with something sufficiently positive afterwards. But its actions for that one day are the result of a series of inputs to a particular algorithm; if we turned off the AI after that day, we would have accomplished moves towards goal G without having to reward its "true" goals at all. And then we could continue the trick the next day with another copy.

For this to fail, it has to be the case that we can create an algorithm which will perform certain actions on certain inputs as long as it isn't turned off afterwards, but that we cannot create an algorithm that does the same thing if it was to be turned off.

Another alternative is to create a superintelligent AI that has goals in a fictional world (such as a game or a reward channel) over which we have control. Then we could trade interventions in the fictional world against advice in the real world towards whichever goals we desire[16].

These two arguments may feel weaker than the ones before: they are tricks that may or may not work, depending on the details of the AI's setup. But to deny the Orthogonality thesis requires not only denying that these tricks would ever work, but denying that any tricks or methods that we (or any human-level AIs) could think up, would ever work at controlling the AIs. We need to assume superintelligent AIs cannot be controlled in any way that anyone could think of.

4.5 In summary

Denying the Orthogonality thesis thus requires that:

1. There are goals G, such that an entity an entity with goal G cannot build a superintelligence with the same goal. This despite the fact that the entity can build a superintelligence, and that a superintelligence with goal G can exist.

[16] Another possibility, for those who believe AIs above a certain level of intelligence must converge in their motivations, is to have a *society* of AIs below this level. If the AIs are closely linked, this could be referred to as a superorganism. Then the whole superorganism could be setup to have any particular goal and yet have high intelligence/efficiency. See http://lesswrong.com/r/discussion/lw/gzl/amending_the_general_pupose_intelligence_arguing/ for more details.

2. Goal G cannot arise accidentally from some other origin, and errors and ambiguities do not significantly broaden the space of possible goals.

3. Oracles and general purpose planners cannot be built. Superintelligent AIs cannot have their planning abilities repurposed.

4. A superintelligence will always be able to trick its overseers, no matter how careful and cunning they are.

5. Though we can create an algorithm that does certain actions if it was not to be turned off after, we cannot create an algorithm that does the same thing if it was to be turned off after.

6. An AI will always come to care intrinsically about things in the real world.

7. No tricks can be thought up to successfully constrain the AI's goals: superintelligent AIs simply cannot be controlled.

5 Conclusion

It is not enough to know that an agent is intelligent (or superintelligent). If we want to know something about its final goals, about the actions it will be willing to undertake to achieve them, and hence its ultimate impact on the world, there are no shortcuts. We have to directly figure out what these goals are (or figure out a way of programming them in), and cannot rely on the agent being moral just because it is superintelligent/superefficient.

6 Acknowledgements

It gives me great pleasure to acknowledge the help and support of Anders Sandberg, Nick Bostrom, Toby Ord, Diego Caleiro, Owain Evans, Daniel Dewey, Eliezer Yudkowsky, Vladimir Slepnev, Viliam Bur, Matt Freeman, Wei Dai, Will Newsome, Paul Crowley, Mao Shan, Alexander Kruel, Steve Rayhawk, Tim Tyler, John Nicholas, Ben Hoskin and Rasmus Eide, as well as those members of the Less Wrong online community going by the names shminux, and Dmytry. The work was funded by the Future of Humanity Institute (FHI), in the Department of Philosophy of Oxford University. The FHI is part of the Oxford Martin School.

7 Notes and References

Adams, E. M. (1980). Gewirth on Reason and Morality. *The Review of Metaphysics, 33*(3), 579-592.

Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines, 22*(4).

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advance Artificial Agents. *Minds and Machines, 22*(2), 71-85.

de Fabrique, N., Romano, S. J., Vecchi, G. M., & van Hasselt, V. B. (2007). Understanding Stockholm Syndrome. *FBI Law Enforcement Bulletin (Law Enforcement Communication Unit), 76*(7), 10-15.

Gewirth, A. (1978). *Reason and Morality.* University of Chicago Press.

Hume, D. (1739). *A Treatise of Human Nature.*

Hutter, M. (2005). Universal algorithmic intelligence: A mathematical top-down approach. In B. Goertzel, & C. Pennachin (Eds.), *Artificial General Intelligence.* Springer-Verlag.

Kornai, A. (2013). Bounding the impact of AGI. *Oxford 2012 Winter Intelligence conference on AGI.* Oxford.

Laplace, P.-S. (1814). *Essai philosophique sur les probabilités.*

Omohundro, S. M. (2008). The Basic AI Drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Artificial General Intelligence: Proceedings of the First AGI Conference* (Vol. 171).

Rawls, J. (1971). *A Theory of Justice* . Harvard University Press.

Sandberg, A., & Bostrom, N. (2008). Whole brain emulation: A roadmap. *Future of Humanity Institute Technical report, 2008-3.*

Schmidhuber, J. (2007). Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial General Intelligence.* Springer.

Wang, P. (2011). The assumptions on knowledge and resources in models of rationality. *International Journal of Machine Consciousness, 3*(1), 193-218.

Waser, M. R. (2008). Discovering the foundations of a universal system of ethics as a road to safe artificial intelligence. *Biologically inspired cognitive architectures: Papers from the AAAI fall symposium*, (pp. 195-200).