Points of failure

Direct specification in AGI alignment

Elias Dokos

Presented to the Faculty of Arts through the
Department of Philosophy in Partial Fulfilment of the
Requirements for the Degree of Master of Research at
Macquarie University

Supervisors: Paul Formosa, Richard Menary

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Elias Dokos

Abstract

Some have critiqued the strategy of explicitly formalising and implementing a value structure in the design of ethical Artificial General Intelligences (AGIs). I build on these critiques by providing a conceptual account of the issues with direct specification, demonstrating its in-principle unviability when compared to implicit and indirect approaches. I begin with a consideration of the factors involved in AGI risk, and the need for risk mitigation. The design of AGIs which are motivated towards ethical behaviour is a key element in risk mitigation. A natural approach to this problem is to directly specify values for the AGI, but this approach necessitates two fatal consequences: an axiological gap between any potential AGIs and humans, and the immutability of this gap. Indirect approaches evade both of these consequences. I construct an account of the axiological gap and argue for its inevitability under direct specification.

Contents

Introduction

As rewarding and useful as AI technologies are their associated risks are many, ranging from illicit human use of these powerful tools to dysfunctional behaviours of the tools themselves. The emerging field of machine ethics (Wallach & Allen, 2010; Anderson & Anderson, 2011; Cave et al, 2018) is the attempt to tackle a facet of the risks involved in creating AI that are, in some sense, able to make ethical decisions. AIs, unlike other tools, are intelligent. This morphs our standard conception of the function of tools into something more generative, where new and unpredictable behaviours are generated by internal cognitive structures. In this vein it seems plausible to conceive of AI as being able to make ethical decisions in some sense; to produce behaviour that arises from moral or quasi-moral values.

The extent to which we take this as an *actual* instantiation of moral behaviour will depend on the nature of the machine itself. It could be argued that, if machines are incapable of thought, then they are incapable of being moral agents (Parthemore & Whitby, 2013). This can be framed as the dichotomy between "weak" and "strong" AI (Russell & Norvig, 2010 pp. 1020-1033). The distinction concerns whether the AI is capable of thought, i.e "strong", or only capable of simulating thought, i.e "weak". Russell and Norvig quote Dijkstra in saying that "*The question of whether Machines Can Think… is about as relevant as the question of whether Submarines Can Swim.*" (Ibid, p. 1021). There is a point here, that the discussion may turn on a definitional point rather than the eminent relevance of the behavioural fact that the submarine is underwater and moving. Whether or not it is *swimming* through the water seems beside the point.

We can, however, take the question of whether machines can think as a question of their internal features which bear relevance to other substantial questions of morality, cognition, and computation. For instance, machine consciousness has a significant bearing on moral questions regarding and relating to machines (Wallach et al, 2011; Torrance, 2008). The related examination of machine moral agency (Sullins, 2011; Moor, 2011) and potential moral responsibility (Gunkel, 2017; Nyholm, 2018; Hakli & Makela, 2019) is central to our

ethical assessment of the uses of machines in medical (Inthorn et al, 2014; Whitby, 2014), military (Sparrow, 2007; Lokhorst & van den Hoven, 2011; Leveringhaus, 2018), transportational (Goodall, 2014; Hevelke & Nida-Rümelin, 2015), and other contexts.

We could also provisionally set aside questions of whether machines can attain a certain level of agential status and, as it were, simply examine the underwater movement of the submarine. In the ethical context this can consist of an engagement with machine ethics as an operation in risk management; the design and construction of machines that behave as we would want. I will be working along these lines. Of particular importance is the construction of ethical Artificial General Intelligence (AGI). AGIs are hypothetical artificial intelligences that can engage in domain-general cognition (Pennachin & Goertzel, 2007 p. 1). They can engage with information in various forms, and solve problems in various spaces, without the need for extensive training in those spaces. In this sense at least, AGI are analogous to human minds. Humans can engage in reasoning over unfamiliar terrain, learning to pick up on relevant features of that terrain, and to ignore irrelevant features. We have our limitations, as do all minds, but it doesn't seem like our limitations are of any in-principle cognitive scope, only cognitive capacity.

The field of AGI safety encompasses a wide range of strategies. These include the design of social policy to guide safe AGI development (Bostrom et al, 2016; Torres, 2018), as well as proposals for constraining AGIs (Armstrong et al, 2012; Bostrom, 2014). AGI alignment is the subfield of AGI safety concerned with aligning AGIs with human values (Yudkowsky, 2016; Bostrom, 2014, Muehlhauser & Bostrom, 2014). There are many technical issues involved in such a project (Amodei et al, 2016; Everitt et al, 2018). There are also many philosophical issues, as everywhere, but one stands out in particular: the problem of determining the ethics the AGI itself should have. I will not be answering this problem. Instead I will argue that it is futile to make such a decision. We should rather look towards allowing the AGI to learn its values using a *process* that we decide on, rather than implementing our preferred values directly. This direct approach is what the philosopher Nick Bostrom calls "direct specification" (Bostrom, 2014; pp. 169-172). Direct specification involves "*trying to explicitly define a set of rules or values that will cause even a free-roaming superintelligent AI to act safely and beneficially*" (p. 169). The alternative is "indirect normativity" (pp. 256-259), where the AGI is built such that "*it can discover an appropriate set of values for itself by reference to some implicitly or indirectly formulated criterion*" (p. 169).

My aim in writing this thesis is to convince the reader that direct specification is an unviable strategy for AGI alignment. In pursuit of this aim I will use conceptual tools from cognitive science, the philosophy of computation, and meta-ethics. Extant critiques of direct

specification emphasise moral disagreement and the historical inadequacies of ethical theory (Bostrom, 2014 pp. 256-259; Yudkowsky, 2004 p. 14; Muehlhauser & Helm, 2013 p. 105). I will make a stronger claim: direct specification necessarily disconnects AGI values from our own and is therefore inadequate for domain-general ethical behaviour. This is to say that the act of direct specification will cause the AGI to have values which are subtly, or markedly, different from our own, and that this difference is a cause of concern.

In the first chapter, *AGI risk and alignment*, I will lay the ground which precedes my argument. A consideration of the risk that AGI poses, in particular the risks associated with artificial superintelligences, frame the importance of this problem. The concept of an "intelligence explosion" (Good, 1965) will be articulated; I will formulate a simple mathematical model of such an explosion, which will allow for the identification of what must be the case for such an explosion to *not* occur. I will also articulate the notion of "convergent instrumental goals" Bostrom (2012) and the analogous concept of "basic AI drives" (Omohundro, 2008), which form a key factor in concerns associated with AGI risk. This will lead to a section on AGI alignment where I will present the dichotomy between direct specification and indirect normativity. Attempts of, and general approaches to, direct specification will be considered, as well as extant critiques, which I will elaborate on.

In the following chapter, *Points of failure*, I will build on extant critiques of direct specification by drawing a conceptual account of the points of failure involved in the process. This account will be used to motivate the stronger claim, that direct specification is in-principle unviable. I will begin by indicating the distinction between instrumental and final values, and I will argue that AGI alignment should aim to be an alignment of final values. I will then illustrate how direct specification involves the translation, or mapping, between frames. The potential attempt to specify our final values directly will be examined, and I will show how it involves at least one, and more likely two steps of abstraction away from our own values. Then I will argue for why this is a problem, and why AGI is better suited to resolving these issues. Finally, I will consider the case where the hierarchical model of evaluative reasoning does not hold, and will defend its employment.

This thesis is part of a wider effort to take seriously the risks posed by Artificial General Intelligence. Mitigating these risks will involve efforts from a variety of specialisations, including but not limited to cognitive science, computer science, mathematics, and philosophy. I have made use of conceptual tools from these areas, particularly where they seemed most instructive, and important in formulating my arguments. One theme which is present throughout this work is a distinction between methods of programming artificial intelligences which, to an extent, mirror the distinction between direct specification and indirect normativity. Ultimately the question of whether direct specification should be

implemented will be answered not only with reference to its ethical viability, but to its pragmatic viability. Where possible I have avoided making claims about normative ethics, because these issues are downstream of the eminent unviability of direct specification.

I: AGI risk and alignment

AGI risk

AGIs have been described as potential "existential risks" (Bostrom, 2014; Barrett & Baum, 2016; Yudkowsky, 2008). Existential risks are those which threaten the survival of the human race itself, or severely impact its potential (Bostrom, 2009; Torres, 2016). There are two key claims that underpin such a weighty description, which can be framed as two questions one might pose to the AGI doomsayer: *how* and *why?* How can an AGI pose such a risk? Why would an AGI pose such a risk? The doomsayer has good reason, I will argue, to claim these questions have answers for a significant slice of the space of possible minds. In other words, there is a not-insignificant chance that an AGI will emerge which has the ability and inclination to pose an existential risk.

The classification of AGI as an existential risk may seem odd, particularly when compared to other existential risks like uncontrolled climate change, global nuclear war, pandemics, and so on. This is a significant claim, but upon examination it has some justification. A core point of concern with AGI stems from its universal capabilities. It is not constrained to a particular domain, and so the consequences of undesirable behaviour are similarly unconstrained.

How?

For something to be an existential risk, it should have the causal ability to threaten the survival of humanity. For the moment, let us consider a subset of the category of AGIs, known as *artificial superintelligences*. A superintelligence[1] is an AGI that has attained a very high level of intelligence, a level that profoundly exceeds our own (Bostrom, 2014 p. 63). I will follow Legg and Hutter's informal definition of intelligence as "an agent's ability to

[1] I will refer to artificial superintelligence as simply superintelligence. This is just a shorthand; I am not implying that any superintelligence must be artificial.

achieve goals in a wide range of environments" (2007, p. 402)[2]. Superintelligences have the ability to bring a wide range of desirable world-states into existence. Their capabilities allow for the enactment of significant local and global changes.

A crucial featurre altering the actual world is the employment of physical manipulators. Yudkowsky (2008) describes one potential method of breaking this barrier where a superintelligence solves the protein folding problem and, through the internet, pays a lab that can synthesise DNA and peptide sequences according to instructions (of which there are plenty), which form a simple biological nanosystem that can manipulate matter to form more complex systems, bootstrapping up to extensive and effective physical manipulators. This is of course just one possibility; it is highly likely that a superintelligence would be able to come up with a better plan. The design and creation of physical manipulators and sub-agents grant the potentially widespread control of physical reality.

A superintelligence would be capable of wreaking havoc, were it so inclined. It is tempting to conceptualise such havoc in terms of actions like comprehensive cyberattacks on governments, military agencies, and corporations; crashing the stock market, causing economic collapse; and mimicking remote controller commands to drones and other technologies. In fact, these are all methods that are probably quite inefficient for a superintelligence; it can achieve faster and more comprehensive results by manipulating matter on the atomic scale. They are also somewhat anthropocentric, which should arouse our suspicions. It is more likely that the havoc wreaked will be on the atomic scale, at a rate too fast for us to comprehend perceptually. "The AI neither hates you, nor loves you, but you are made out of atoms that it can use for something else." (Yudkowsky, 2008 p. 333). And such is the trap of anthropocentrising AGI risk: if we think about AGI risk in manifestly human terms, we are probably not going to capture the actions that a superintelligence might actually carry out. Suffice it to say that more intelligence means more and more effective and efficient means of achieving goals, including means that are incomprehensible to us at our current level of scientific understanding and intellectual capacity.

An issue might be posed here: maybe the existence of a superintelligence would be dangerous, but it's a large leap from mere AGI to superintelligence; a leap that we are likely

[2] Bostrom distinguishes between different kinds of superintelligence. One of them, *collective superintelligence* (2014, pp. 65-68) is a superintelligence that is the emergent result of different intelligences. It could be a large community of brilliant scientists working in tandem, organised in a system that we have not yet managed to create, or a collective system of different artificial intelligences. It could be argued that such a superintelligence does not conform to this definition of intelligence. To this I will simply say that, as long as this collective superintelligence is goal-directed, its system-wide intelligence can still be measured according to its ability to achieve its goals in a wide range. However, collective superintelligence may involve a significantly different approach to alignment; here I will concern myself with superintelligent *agents* only.

incapable of making. This may be the case, but superintelligence does not need to be directly created. All that is needed is an AGI that has reached a critical point of intelligence: a point where it is capable of self-improvement, or the design of other, superior, AGIs. From this point an explosion of intelligence will likely occur.

First conceived of in serious academic circles by Good (1966), the concept of an "intelligence explosion" is a hypothesised more-than-linear growth of an artificial agent's intelligence once they reach the intellectual capacity where they can design and implement better versions of themselves. The story goes like this: we eventually develop an AI intelligent enough to upgrade itself, which it does. So now we have an improved AI. This new AI is better at upgrading itself than the original. When it upgrades itself, the next iteration will be a yet greater improvement. And so on, and so on, until eventually (or not so eventually) we have a superintelligence. Intelligence explosion, and indeed any incremental cognitive self-improvement or external agent design that doesn't plateau at some pre-superintelligent point accounts for the possibility of superintelligence without our having the ability to directly design one.

There are several points in the "how" argument that could be challenged. To highlight this, I will frame it in the form provided to us by Chalmers (2010 p. 6):

1. There will be AI (before long, absent defeaters).

2. If there is AI, there will be AI+ (soon after, absent defeaters).

3. If there is AI+, there will be AI++ (soon after, absent defeaters).

4. There will be AI++ (*i.e superintelligence*—soon after, absent defeaters).

Chalmers uses the term AI rather than AGI, but he means AI in the narrower sense of a human-equivalent artificial mind. Perhaps this could be made even narrower: a mind that is capable of self-improvement. We might say that human minds are not at this level, but I am not so sure. We are limited by biology: if a human-equivalent mind was instantiated in digital nodes, with the consequent ability to add more nodes and networks with relative ease, self-improvement could be as simple as stumbling on an addition which is beneficial. It is likely that we would not be so blasé about it, nor would an AGI, but at the very least the level of intelligence required to improve upon a digital or electronic mind is lower than that

required to improve upon a biological mind. Consider the progress humans have made in designing and improving digital minds, compared to biological minds.

It is instructive to consider artificial minds as a subset of the much larger set of *possible minds* (Yudkowsky, 2006). Here we can ask two sets of two questions: do AGIs exist in the set of possible minds? Can we find them in that set? And: does superintelligence exist in the set of possible minds? Can it be found in that set?

The possibility of AGI

There seems to be no good reason why general intelligence is *in principle* restricted to the biological domain. Many arguments against AGI tend to make slightly different claims. For instance, as Chalmers (Ibid pp. 8-11) notes, Dreyfus (1972), and Penrose (1994) argue that machines—meaning classical computer systems—can never reach the level of cognitive capacity reached by humans. In critiquing optimism about artificial cognition Dreyfus writes: "*Underlying their optimism is the conviction that human information processing must proceed by discrete steps like those of a digital computer, and, since nature has produced intelligent behavior with this form of processing, proper programming should be able to elicit such behavior from digital machines, either by imitating nature or by out-programming her.*" (1972 p. 67). Bringsjord et al (2012) make an opposite move which falls into the same trap when they argue that it can be shown mathematically to be the case that humans can never design an AI as intelligent as ourselves. This is a result that emerges from modelling both human and artificial intelligence as Turing machines (pp. 403-404).

Before examining this issue, I would note that Bringsjord and colleagues' conception of AGI design as an agent designing a smarter (or equivalently smart) agent is not quite accurate in the human case. AGI design, as with most big projects, employs teams of agents. They incrementally build on the advances of past generations, and allocate sub-problems to other agents and teams. Even if it were the case that an agent couldn't design a smarter agent, a collective system of distributed cognition (Mangalaraj et al, 2014) could well design an agent more intelligent than any given member.

There is a common thread to Dreyfus and Penrose's approaches, and Bringsjord and colleagues' approach. These approaches each narrow the space of possible minds in particular ways. Dreyfus (1972) and Penrose (1994)[3] narrow the space of artificial minds into Turing-machine equivalent mind designs, i.e discrete steps drawing from a finite set of rules.

[3] "*The important thing for our purposes here is that both top-down and bottom-up computational procedures are things that can be put on a general-purpose computer and are therefore both to be included under the heading of what I am referring to as computational and algorithmic.*" (Penrose, 1994 pp. 18-19)

Bringsjord and colleagues narrow the space of *both* human and artificial minds into Turing-equivalence. Turing equivalence is a system's computational equivalence to a classical Turing machine: a machine that is comprised of an infinitely long tape upon which discrete operations are carried out, with reference to a finite set of rules (Galton, 2006).

There is good reason to suppose that human minds are not Turing machines (King, 1996; Goldin & Wegner, 2008; Copeland & Sylvan, 1999). If this claim is true, then the space of possible *artificial* minds is likely not constrained to Turing machines. "Soft computing" (Bonissone 2010) provides a radically different paradigm of agent design where engineers provide broad specifications of architectures, goals, and hardware, and allow the systems to self-develop. Under these soft computing paradigms engineers do not exhaustively specify the functional characteristics of the system. This avoids the pitfalls of the kind that Dreyfus identifies (1972 pp. 168-183), where explicit and exceptionless rules and relations are insufficient for generalised and contextual behaviour.

All this said, it would be wrong to entirely dismiss the possibility of some developmental obstacle that we cannot surmount, or our techniques of AI engineering are unsalvageably ill-suited to the task of AGI design. The possibility of such an obstacle exists, but as a mere possibility it is not sufficient to delegitimise AGI safety efforts. Here, as with all objections to AGI risk, we should be aware that we are in the game of risk management, which entails probabilistic reasoning. It is not enough to say, for instance, that it is unlikely that my house will catch on fire. This may be so, and we can all hope that it is unlikely, but it seems to be a poor argument against taking out insurance. Even more so when the risk is on the existential level, not just the structural. If we can be certain, or very close to certain, that AGI is impossible, or that some necessary step in the arguments for AGI risk is very close to certainly wrong, then that is a stronger reason to dismiss the AGI risk hypothesis. If, however, there is enough reason to give some not-insignificant level of credence to the argument, then the possibility becomes a risk worth taking seriously.

The possibility of superintelligence

Another point at which the intelligence explosion hypothesis could be challenged is the model of an iterative process of self-improvement. Regarding the rate of self-improvement, Bostrom (2014 p. 80-94) defines two key variables: optimisation power and recalcitrance. Optimisation power is a measure of the effort that is turned to the purpose of optimisation, and recalcitrance is the system's lack of responsiveness to optimisation. As intelligence increases so does optimisation power, but the rate at which the recalcitrance of the agent increases is key to the question of whether there will be an intelligence explosion. It is reasonable to suspect that higher levels of cognitive capacity are also more difficult to

optimise, even with increasingly clever optimisation strategies. The central question, however, is how this difficulty scales with ability.

Under the reasonable assumption that optimisation power scales with intelligence, explosion occurs if $\sum_{n=1}^{\infty} \frac{S_n - S_{(n-1)}}{r(S_n)}$ diverges, where S_n is the intelligence of the system at the nth iteration of improvement, and $r(S_n)$ is the recalcitrance of that system. That is to say, at each iteration of self-improvement some intelligence is "added" to the system (for modelling purposes, this is the case whenever we can say that a system S_n is more intelligent than some other system S_{n-1}, otherwise we are not talking about improvement). If the sum diverges, the total intelligence of the system can reach an arbitrarily high point (mediated by the recalcitrance at each step) after an arbitrary number of iterations of self-improvement. If the sum converges, then there is a limit to self-development which can be reached quite early in the process, for practical purposes. It is vital to note that this limit, if it exists, may be quite high relative to human capacities, so this does not necessarily preclude the possibility of a superintelligence.

If recalcitrance is linearly related to system intelligence (i.e they both scale at a consistent rate relative to one another) then intelligence growth will itself be linear, in lieu of other relevant factors. If the rate of change of recalcitrance is higher in *degree* than the rate of change of optimisation power then intelligence growth will plateau, and if the rate of change of recalcitrance is lower in degree than the rate of change of optimisation power than intelligence explosion will occur. This is to say: under this model explosion is contingent upon the highest exponent of the recalcitrance r compared to intelligence S_n. This means that there is a possibility that intelligence explosion will not occur, though this applies to qualitative cognitive developments, (Bostrom, 2014) and not to other, much easier, forms of self-improvement like simply adding more memory to the system.

Yudkowsky (2008) refers to the rate of self-improvement with the simple model of the number of self-improvements which, on average, result from a single act of self-improvement. If each act of self-improvement results in less than one average consequent self-improvement, then intelligence will plateau. If, on average, each act of self-improvement results in greater than one consequent self-improvement, then intelligence will explode.

This is, however, a much stronger claim than is needed for superintelligence. All that is needed for the emergence of superintelligence is for S_n to have a lower bound that is on the level of superintelligence.[4] In other words: it is very unlikely that we can make claims about the relation of recalcitrance to optimisation power for *any* arbitrarily high level of

[4] More pragmatically: at the level where it poses an existential risk.

intelligence. However, all that is needed for superintelligence to be possible via the route of intelligence explosion is for this relationship to hold *at least* up until the point where superintelligence is reached.

Why?

Even if the existentially risky capabilities of AGI were granted, the question of why it would use those capabilities in an existentially risky way remains. This is the core concern behind the project of AI alignment. The central factor of dangerous AGI motivation is not maliciousness, but indifference towards humans and human interests. An AGI will be inclined to achieve its goals in the most effective way possible. If its goals do not align with ours in a suitable way, its actions will not align with our wishes, nor with our flourishing as a species.

The core claim in answering the *why* question is that AGIs have a high chance of inevitably having certain specific goals, and that those goals do not include a concern for humanity. Omohundro (2008) describes several "basic AI drives"[5], which are the natural result of goal directed systems. These include self-improvement; the preservation of final goals; self-protectiveness; resource acquisition; and others. The reason these are basic drives is that they allow for the improved attainment of almost any fundamental goal. Bostrom (2012) expands on Omohundro's approach by explicitly differentiating between instrumental and final goals, where final goals are the attainment of the core values of an agent, and instrumental goals are those which are adopted because they help achieve the agent's final goals. He then describes several "convergent instrumental goals", which are goals which will be converged on by potential agents with a wide range of final goals. The distinction between instrumental and final goals is one which will play a significant role in the following chapter.

The account expressed by Omohundro and Bostrom is plausible. If there are actions and attitudes which are generalisable in their utility, i.e which are useful in a wide range of situations, then we can imagine that agents who are good at achieving their goals will tend to adopt them. We should expect a rational superintelligence with some arbitrary final goal to converge on certain instrumental goals. Exceptions would involve cases where the final goal is itself contradictory to an instrumental goal. For instance, if an AGI were built with a final

[5] Bostrom (2012) remarks that the term "AI drive" has unhelpful connotations, and that "[o]*ne would not normally say that a typical human being has a "drive" to fill out their tax return, even though filing taxes may be a fairly convergent instrumental goal for humans in contemporary societies (a goal whose realization averts trouble that would prevent us from realizing many of our final goals).*" (p. 76)

goal that inclined it to avoid collecting resources, or a final goal to remain at a certain level of intelligence, then it would not converge on resource acquisition and cognitive self-improvement.

Resource acquisition is an example of an instrumental goal that poses an existential risk. Bostrom (2003) introduced the thought experiment of a "paperclip maximiser": an AGI constructed by a paperclip company with the final goal to maximise the number of paperclips. Such an AGI may begin by designing more and more efficient methods of paperclip production, but humanity is unfortunately not very devoted to paperclip production, which represent a miniscule fraction of the global economy. At some point it would serve the AGI's final goal to acquire more resources than we might be comfortable with, in order to make more paperclips. At some point the AGI will want to convert all matter on the planet into paperclips, and if able, it will want to tile the entire galaxy with them. If the alternative is a less paperclippy world, then given the ability to do so a paperclip maximiser will be inclined to act towards that end.

Waser (2008) agrees with the idea of basic AI drives, or convergent instrumental goals, but argues that "*ethics is actually an attractor in the state space of intelligent behavior*" (p. 1). Evolution and artificial life simulations, he reasons, have shown that cooperation and sociality are instrumentally beneficial, and that selfishness can be instrumentally harmful. This is also supported by game-theoretic models: selfishness is a rational strategy when a particular interaction is only one-off, but altruism is a superior strategy when interactions are reoccurring. All this is indeed true *for humans*. The key factor that determines whether cooperation is beneficial or not is whether the benefits of cooperation outweigh the downsides, which itself is a function of ability. Risks of defection from human cooperation, or otherwise problematic AGI behaviour, become significant once the process of self-improvement leads to the agent surpassing a stage of competence where cooperation is beneficial (Shulman, 2010). For instance, humans do not cooperate with many other organisms. Where it serves our interests, and because we have the ability to do so, we are willing to kill, to harvest, and to otherwise further our own interests at the expense of the potential interests of other organisms. This is not a universal description of our relationship with non-human organisms, but it is also clear that we do not exist in a state of mutual collaboration with many organisms; the relationship is more one-sided. The fact that we have the ability to act in our own interests, rather than being forced to cooperate from the inability to be entirely self-interested, is a factor in this one-sided relationship. There are other factors alongside ability, but the ability to do something is a necessary condition for something to be done.

Moral considerations can serve as both benefit and barrier to non-moral goal completion. They entail the proscription of certain actions and methods to obtain one's goals, which can in turn cause others to act in support of those goals. It would make less sense for me to cooperate with a mosquito than for me to cooperate with another human. Assuming I had no ethical drive, the instrumental advantages of cooperation and the disadvantages of defection are entirely different in these two cases. Defection in human interactions can result in consequences ranging from the social to the punitive. The question we should ask is: what is holding a superintelligence to the same instrumental standard? This applies to both the consequences of defection, and the trade-off between a superintelligence acting in its own interests against reaping whatever benefits we might grant it for cooperating. Instead of hoping for the inevitability of pro-moral and humanitarian inclinations, it is vital that we attempt to build AGI with such inclinations as part, or all, of its core values.

Of course, not everyone will agree with this position. Waser makes the argument that ethics is a convergent instrumental goal, appealing to the core self-interest of goal-directed agents and asserting the convergent status of pro-moral behaviour with reference to this goal. But others may wish to make a stronger claim. Kant wrote that "*…the fitness of the maxim of the good will to make itself into a universal law is itself the sole law that the will of* ***every*** *rational being imposes upon itself, without underpinning it with any incentive or interest as its foundation.*" (GMM 4:444, 2012 p. 55, emphasis mine). In his consideration of the Kantian perspective of AGI alignment Chalmers (2010) notes the possibility that Kant is employing a particular view of rationality which is already "infused with morality" (pp. 28-29).

Be this as it may, it could be argued that this element of AGI risk presupposes that morality is not a convergent structure in the space of rational or empirical reason. I suppose it does, but the person who is an AGI risk sceptic for this reason would need to fulfil two stringent criteria. The first is a confidence that whatever kind of rationality an AGI has will conflate their potential acceptance of moral facts with their motivation to act morally, i.e *internalism* about moral motivation (Dancy, 1993 p. 1). The second is a level of credence in *both* their views about the rational inevitability of a certain set of moral attitudes and their internalist views. If there is room for doubt, if the credence is sufficiently imperfect, then the risk remains, though no doubt lessened in probability.

AGI Alignment

AGI alignment is, loosely speaking, the project of building "friendly AGIs" (Yudkowsky, 2004; Taylor et al, 2006). The idea of a friendly AGI is self-explanatory enough, as long as we keep it somewhat vague. Let us remain vague for the moment, to identify what most would probably agree with. In this vein we could say that we—*sort of*—know what we want. At any rate, we certainly know what we don't want, which includes a large set of outcomes like the galaxy being tiled with paperclips, or more visceral, if naive, images of robot hordes marching through the streets shooting laser beams at anything with a heartbeat. It is safe to say that, at the very least, we can agree that we are in the game of avoiding such outcomes, though there will be disagreements at the edges of this set. Maybe some would argue that an eternity of humans locked in Nozick's (1974) experience machine would be acceptable, while most would disagree. But on the whole, the set of outcomes that almost everyone would agree are undesirable is large enough.

It is likely that the most effective, though perhaps the most difficult, path to avoiding what we don't want is to determine the AGI's motivations in such a way that it is disinclined to do what we don't want it to do, and inclined to do what we want. Other possibilities include attempting to constrain the AGI in some way, to restrict its ability to act in the world and consequently its capacity to act in ways we do not want (Bostrom, 2014; Armstrong et al, 2012). These methods are generally compatible with AGI alignment, and we should employ them where possible, but we should not make do without alignment. The reason why is quite simple: an AGI will have goals,[6] and it will be motivated to, and potentially quite capable of, achieving these goals. External constraints are barriers to goal completion, and it will be motivated to overcome them. It may or may not be capable of doing so, but it will *want* to, and will likely put its best efforts into doing so. If we ensure that it is motivated from the outset to act in ways we want, then an AGI that is better at being successful will work in our favour, not against it. That said, we can and should use all our cards in this matter, where there are no conflicts and incompatibilities between methods.

A central factor in this approach to AGI alignment is the question of what, exactly, the AGI's motivations should be. Above, I loosely described the goal of AGI alignment as the creation of a friendly AGI. A more specific goal is the creation of an AGI that is aligned with our values. Insofar as we can describe an AGI as having values, we can identify whether those values are aligned with our own. If we succeed in this alignment, issues of problematic AGI behaviours will be resolved, at least to a significant degree. Immediately a concern arises, namely: what does it mean for AGI values to be aligned with our values? This concern will be considered at length in the following chapter.

[6] I am doubtful that an intelligence could exist that does not have goals. If such an intelligence could exist, then motivation selection would not apply, but it is likely that neither would risk.

Nick Bostrom draws a distinction between two approaches to value selection: *direct specification* and *indirect normativity* (2014, pp. 169-173). Direct specification describes the solutions to value specification that involve giving the AGI explicitly formulated values that will serve as the final motivational structure of the AGI in all situations. Indirect normativity is the converse approach, which allows the AGI to learn the right values by instead specifying a *process* for value acquisition rather than the values themselves.

Direct Specification

The most intuitively obvious method for solving this issue is simply to specify rules or values. The central idea is to construct a motivational structure for the AGI that will generate ethical actions in any context. In part, this method is made more appealing by the nature of contemporary narrow artificial intelligences, many of which involve the need to specify cost functions and other algorithms for aligning AI outputs with what we prefer. Direct specification is abundant in machine ethics, which is likely due to its similarity to ethical research in general, which consists in the attempt to find principles from which all ethical evaluation can be sourced (Bringsjord et al, 2006 p. 39)[7].

There are, broadly speaking, two families of approaches to direct specification: the utilitarian and the deontological. Each has its own intuitive appeal. The utilitarian approach is made more appealing by the nature of extant "soft computing" systems, most notable Reinforcement Learning (RL). Many significant contemporary successes achieved by AI research have emerged from applications of RL. Essentially, this frames an artificial intelligence as part of a Markov Decision Process, which is a system comprised of four sets, (S, A, P_a, R_a): States, Actions, the Probability of an action A in a state S causing a state S', and the Reward for transitioning from S to S' (Wu et al, 2014). An AI created under an RL paradigm attempts to maximise reward by acting within certain states such that more rewarding states are reached. Actions could be moves within a chess game, the generation of words, an embodied movement, and so on.

The concept of value seems to fit naturally into this model, specifically as the *reward function* of an RL agent: the algorithm that determines reward on the basis of observation. Alternatively, it could be understood as a utility function, which motivates in virtue of the expected reward of a particular state of affairs. Utility functions have been widely applied within decision-theoretic AI research (Heckerman, 2013; Chajewska et al, 2000; Soares &

[7] This is not true of all ethical research; the particularist pushback to this view will be examined in chapter 2.

Fallenstein, 2015). It should be noted that the connection from utility-functional approaches to ethical utilitarianism is not one of equivalence; it is simply that utilitarianism bears an aesthetic resemblance to utility function approaches, which can potentially bias programmers towards such ethical approaches.

The proposal space of *aligned* AGI utility functions is mostly unpopulated, in part because such a utility function would represent a major step forward in AGI design that has not yet been made. Hutter's AIXI model (2007) is a proposed universal intelligence which is a mathematical formalisation of an epistemic strategy for choosing hypotheses about its environment given a reward function, but it does not specify the reward function itself; the problem of defining the right reward function for an AGI remains.

Considerations of broader reinforcement learning strategies for ethical AGI are present in the literature. Inverse Reinforcement Learning (IRL) is an approach that attempts to infer the utility functions of other agents, such as humans (Natarajan et al, 2010; Kuleshov & Schrijvers, 2015; Hadfield-Menell et al, 2016). Given a Markov Decision Process (S, A, P_a, R_a), an outside observer has access to the set of states, actions, and probabilities. IRL is the process of inferring the agent's reward function from observations about the other three sets. An appropriate utility function for an artificial agent can be constructed from knowledge of human utility functions, and knowledge of human preferences can be formalised with their utility functions.

There are many ways to conceptualise and implement ethical frameworks in artificial intelligences with utility functions. Utility functions are analogous to a source of motivation, but this does not need to translate to an explicitly utilitarian ethical framework implemented by the agent. Take, for instance, two utility functions involving paperclips. One might simply track reward alongside the sheer number of paperclips, meaning the agent seeks to maximise paperclips, and consequently to implement paperclip production on a massive scale. The second utility function might look something like: $U = \begin{Bmatrix} 0\ if\ |paperclips| \neq 200000 \\ 1\ if\ |paperclips| = 200000 \end{Bmatrix}$. In this case, the agent will not maximise paperclips, indeed it will do anything within its power to prevent the creation or destruction of paperclips once there are 200000 paperclips. While obviously not a reasonable example of a utility function we might wish to implement, U is an example of a utility function that implements a rule: THOU SHALT MAINTAIN THE NUMBER OF PAPERCLIPS AT 200000, or equivalently, THOU SHALT NOT EXCEED OR FALL BELOW 200000 PAPERCLIPS.

Eckersley (2019) has criticised IRL, along with all utilitarian solutions to the alignment problem. He references an impossibility theorem in population ethics put forward by Arrhenius (2000) which shows that solutions to certain unfavourable consequences of

utilitarian theories, such as the “mere addition paradox”, are mathematically incompatible with solutions to other unfavourable consequences. Eckersley applies this to the context of an AGI, countering several potential responses to the impossibility theorem, and extending the proof to indirectly normative objective functions over populations. He demonstrates that there is no simultaneously compatible set of constraints subject to a cyclic impossibility theorem that can all be below a particular lower bound of normative uncertainty.

Deontological approaches are appealing to some because of their ability to be formalised in deontic logic. Deontic logic is an approach to ethical modelling which deals with normative relations (Bringsjord et al, 2006). Hooker and Kim (2018) formalise several deontological principles, beginning with the generalisation principle - that “*the reasons for my action are consistent with the assumption that everyone with the same reasons takes the same action*” (Ibid p. 4). The formalisation of the generalisation principle then leads to other principles. Arkoudas, Bringsjord, and Bello formulate an automated system to determine whether a given action comports with deductive validity within some given “an expressive deontic logic *L* of high practical relevance, and an efficient algorithm for determining theorem-hood in *L*” (2005 p. 2).

This approach naturally lends itself towards Good Old-Fashioned AI (GOFAI) approaches: the attempt to design artificial intelligences in the form of Turing machines, sequential and algorithmic and symbolic processing (Davenport, 2013 p. 44). This stands in contrast to the soft computing approach which aligns itself more naturally with utilitarian style specification, based in connectionist architectures. This is not to say that these relations are relations of equivalence. Utilitarian value specifications can be (or so it is hypothesised) instantiated with deontic logic (Hooker & Kim, 2018 p. 5), and deontological value specifications can be instantiated in reward functions by simply defining negative utility in terms of failure to comply with a rule or rules, as described above.

One approach to direct specification is to take an extreme GOFAI approach and algorithmically specify all the normative relations that exist: all the relationships between actions, or world-states, and moral value or desirability. An example is illustrated in Horgan & Timmons' account of an epistemic normativity profile: "*an enormous list of completely specific cognitive transitions from one specific potential* [total cognitive state] *to another that a given agent ought epistemically to make…*" (2009 p. 38). They argue convincingly that this is not possible, due to the frame problem.

The frame problem emerged out of issues in classical AI, which was characterised by its attempts to create artificial agents through systematic axioms and first-order logical propositions, otherwise known as GOFAI. The frame problem was the problem of how such a

system of propositions could be defined without including all the myriad, indeed infinite, factors and non-factors that relate to the problems being considered (McCarthy & Hayes, 1969). The frame problem concerns the specification of causes and effects in first-order logic for some domain. For each effect of an action, there are myriad non-effects of that action. It would be logically consistent, within the system of specified causes and effects, for a non-effect that is *unspecified* by the system to actually occur. To get around this difficulty, axioms for each non-effect will need to be specified, known as "frame axioms" (McCarthy & Hayes, 1969). When the required domain of operations is expanded, the number of necessary frame axioms explodes.

Say that we implemented a directly specified normative framework on an agent that is not *as* hamstrung by the frame problem—a connectionist neural network, for instance (Clark, 2002). What might a nuanced approach to direct specification look like? The idea of an exhaustive list of every possible evaluative stance is one that almost nobody would find epistemically plausible. A more ostensibly attainable attempt to specify an ethical framework would likely involve a smaller set of high-level principles which can be robustly particularised in any given situation. A proponent of direct specification could argue that we can extend over all (or an adequately large subset) of moral decision-making with a limited set of sufficiently abstract principles, or perhaps one super-principle. This may or may not be the case, but the most significant obstacle is their ability to *find* and *frame* those principles. I will examine this in more detail in the following chapter.

Extant critiques of direct specification

Even before Bostrom gave it a term, direct specification has been considered and rejected by several researchers. Eliezer Yudkowsky writes: *"Everyone has their brilliant idea for the Four Great Moral Principles That Are All We Need To Program Into AIs, and not one says, "But wait, what if I got the Four Great Moral Principles wrong?" They don't think of writing any escape clause, any emergency exit if the programmers made the wrong decision. They don't wonder if the original programmers of the AI might not be the wisest members of the human species; or if even the wisest human-level minds might flunk the test; or if humankind might outgrow the programmers' brilliantly insightful moral philosophy a few million years hence. But most who ponder AI morality walk straight into the whirling razor blades, without the slightest hint of fear."* (2004, p. 14).

Muehlhauser and Helm (2013) point out that widespread ethical disagreement, even[8] among experts, should foster suspicion of proposed final solutions to the problem of superintelligence value specification. They pump this intuition with the thought experiment of a "Golem Genie". This is a hypothetical godlike being who has both the ultimate power to shape reality, and a hyper-literalness which will follow a specified moral code to the letter, without the implicit interpretative nuance of, say, a human following instructions (pp. 105-106). The prospect of providing this Golem Genie with a fixed code of conduct is something that should inspire extreme caution and apprehension in anybody with even a modicum of epistemic humility. Bogosian (2017) argues along a similar line: ethical disagreement entails that we should avoid implementing moral certainty into intelligent agents. He points to the intractability of some disagreements; given the assumptions and foundations of different ethical positions, there is no "*halfway point where the competing principles can meet*" (p. 594) for certain disagreements.

The 2009 PhilPapers survey found that, of the 931 professional philosophers surveyed, there was a relatively even spread between the major normative ethical frameworks: 241 accepted or leaned towards deontology, 220 accepted or leaned towards consequentialism, 169 accepted or leaned towards virtue ethics, and 301 had some "other" normative perspective (Bourget & Chalmers, 2014). These results do not even touch on the disagreement *within* those frameworks. Whatever one's perspective on the viability of direct specification, it is a fact that the values that would be specified by one researcher would likely be quite different from those specified by another. Though widespread moral disagreement does not in and of itself prove anything about the in-principle viability of direct specification, it does lend credence to these arguments

These arguments against direct specification I will call *humility arguments*. They turn on the core claim that we are not up to the task of specifying a satisfactory value system for a superintelligence. The humility argument is strengthened by the fact that we have a much higher standard of success in AGI alignment. This is the case for two reasons. First, we are dealing with a potential superintelligence, which is the most significant source of risk that we are trying to avert. Second, we are dealing with the need to be as precise as possible. The far lower standard of success for specifying human values and ethical frameworks has arguably not been reached. Further, it is likely that we have yet to specify one which is satisfactory to even one person without the need for interpretive nuance, often conceptualised as judgement. Judgement, in the sense I am using here, is "*an aptitude for assessing, evaluating, and choosing in the absence of certainties or principles that dictate or generate right answers*" (Thiele, 2006 p. 5). Judgement in this sense is the opposite of the Golem

[8] Perhaps especially.

Genie's literalness, which is the exceptionless algorithmic application of the principles that it has been given, not necessarily interpreted in the way we might wish.

While direct specification allows us to remain relatively confident that we will get precisely what we asked for, this does not mean that we will get precisely what we *want*. Every attempt to comprehensively capture ethical decision-making in a satisfactory way runs an extremely high risk of being either inconsistent, limited, or having consequences that we would not approve of, were they brought to our attention. This may not be much of an issue for narrow AIs like self-driving cars, but the problem becomes apparent when the objects of consideration are AGIs, and particularly superintelligent AGIs. These will certainly experience an understanding that is beyond our grasp, such that we may not have the conceptual tools to adequately analyse them. Not simply unforeseen circumstances, but unforeseeable circumstances.

Donald Rumsfeld famously drew attention to differences in the lack of knowledge: the "known unknowns" and the "unknown unknowns".[9] Any systematic moral framework that purports to exhaust the space of relevant moral considerations with values, action policies, or constraints is at best exhausting the space of known moral considerations. In ethical terms, known unknowns could refer to any number of moral considerations that take place over issues we have not yet experienced, that we *can* apply prior considerations to. Doing no harm, respecting preferences within reason, and so on, are principles we can apply to many new and strange moral encounters.

Unknown unknowns can refer to ethical considerations for which we do not have the conceptual tools to express or understand, let alone resolve. The probability of future issues that are unknown unknowns to us should be taken seriously, as should the probability that our current candidates for value specification will fail to robustly hold in entirely unknown spaces. Often, perhaps always, essential ethical concerns are only coherent once the distinguishing concepts are made explicit. It is likely that many of the concerns of humanity in the near future will seem quite alien to us, and many of the concerns of humanity in the far future will be unintelligible. This becomes even more stark with a superintelligent mind; such a mind is likely to work on issues that we will never have the capacity to understand, even in principle. All attempts at value selection, and all attempts at systemising in general, suffer from the problem of unknown unknowns. With AI alignment the risk of falling prey to this becomes quite significant, assuming that "takeoff" will occur.

[9] https://www.nato.int/docu/speech/2002/s020606g.htm

Indirect Normativity

If I ask my friend to fetch me a beer from the fridge, I can trust that he will fulfil this request without leaving the fridge door open or throwing the beer at my face. This is because I can leave certain preferences of mine underspecified; I trust that my friend has a sufficient grasp of what I would not want, and what he himself would not want. Ideally, I should like to have a similar level of trust in an AGI. Indirect normativity is one way we might move towards implementing such a trust. It describes the approach to motivation selection that takes the route of implicitly or indirectly specifying a goal, process, or policy which might attain a preferable ethical result in that it allows the AGI to learn the right values and methods.

While Bostrom coined the term "indirect normativity", Yudkowsky has described a similar tactic. He describes a proposal for AGI alignment as an "initial dynamic", which will be able to improve itself over time, a dynamic which contains the motivations for such self-adjustment (2004 pp. 9-10). Bostrom motivates indirect normativity with the *principle of epistemic deference* (2014, pp. 258-259), which acknowledges that a future superintelligence will have beliefs that are more likely to be true, and so we should defer to its judgement where possible. This includes allowing an AGI to not have fixed values, and to engage in a learning process which will be superior to ours at a certain point of self-improvement.

Bostrom suggests two potential examples of indirect normativity: Do What I Mean (DWIM) (2014 pp. 270-271), and Moral Rightness (MR) (pp. 266-268). DWIM is the counter to the eternally recurring problem of programming, which is that technology tends to *do what we say*. Computers are very good at competently carrying out unwanted tasks, the results of our failures of specification. But when we instruct other humans, unless they are otherwise motivated, they will tend to interpret the instructions in a way that at least attempts to capture their intended meaning. When my friend grabs a beer from the fridge he is Doing What I Mean, though this is obviously a very different case to AGI alignment. My friend already has a value system in place which proscribes many undesirable interpretations of my request. With an AGI, we would need to define a goal from the outset which allowed it to engage, or learn to engage, in a process of charitable interpretation of either value specifications or of explicit instructions.

MR is the approach to value specification where the AGI is instructed (for some sense of "instructed") to discover and implement the objectively correct values. Of course, there is the possibility that there are no "correct" values to speak of. To account for this possibility, MR includes the specification that, if the search for the correct values turns up empty, then another supergoal proposal be implemented by the AGI. One such proposal is Yudkowsky's

(2004) Coherent Extrapolated Volition (CEV), which is based on a conservative version of ideal observer theory, which is a philosophical approach in ethics, and other domains, based on the idea of defining normative properties and "good" moral judgements in terms of those of a hypothetical ideal observer (Kawall, 2006). The conservative version of this is the concept of "volition", which is essentially an agent's wish if they were better, i.e if "*we knew more, thought faster, were more the people we wished we were, had grown up farther together*" (2004 p. 6).[10] He then outlines the first approximation of a proposal for the *coherence* of the volitions of humanity at large, which will provide at least an initial dynamic for a goal we might wish to see achieved.

A more technical account is Christiano's *A formalization of indirect normativity* (2012), which should be taken as a proposal for a particular instantiation of indirect normativity, not as a formalisation of the entire category of methods. His tentative[11] suggestion is that the AGI engages in an extensive simulation, one which would only be possible once superintelligence has been reached. This involves simulating a person's utility function in a very specific context: an environment with a community of clones of themselves, and access to an idealised computer with limitless computational resources. This proposal may seem odd, but it should be considered alongside volition and other ideal observer-esque proposals. This particular proposal has the advantage of being formalisable: if a mathematical model of a human brain can be generated by a superintelligence, and unbounded computational resources can be simulated as an approximation, then we have a way of formally cashing out an ideal observer situation.

Dewey (2011) proposes the idea of *agent implementation*, which is a process or structure that itself produces reinforcement learning agents with their own reward functions. This agent implementation can design and refine new agents, updating its specification of their reward functions. He claims this avoids the problems of reinforcement learning, where the agent is ultimately incentivised to maximise reward itself. Reward functions are a means to this end. If successful, agent implementation can be designed such that it can iteratively improve over reward function design to better capture human values. This process is achieved through implementing an *uncertainty* about the utility function of the agent implementation itself, and a specification about what kind of evidence it should accept to

[10] In introducing Yudkowsky's proposal, Bostrom (2014) notes that it diverges from ideal observer theory in the sense that it says nothing about what our actual judgements ought to reflect; it is simply a description of a different sense of "will" or "desire" that may be useful in the context of an AGI deciding what we want.

[11] Christiano does not *endorse* this approach, he merely suggests a formalisation: "*I don't think* [indirect normativity is] *an adequate solution to the AI control problem; but to my knowledge it was the first precise specification of a goal that meets the "not terrible" bar, i.e. which does not lead to terrible consequences if pursued without any caveats or restrictions.*" (2012).

update its uncertainty. This uncertainty about utility functions is one way of implementing a capacity to modify goals and values, but Soares et al (2015) believe that it is insufficient: implementing precise utility functions with uncertainty attached involves precisely specifying a function which itself determines the credence the agent would give to a particular utility function. The agent would then be resistant to modification of this uncertainty-determining function itself. They define "corrigibility" as the problem of designing agents that are tolerant of *all* errors in specification.

Another example of indirect normativity is a subset of *bottom-up* approaches to AGI alignment. Wallach and Allen (2009) differentiate between two broad approaches to alignment: top-down and bottom-up, though they do not constrain this distinction to AGIs in particular. Top-down alignment is the algorithmic formalisation of action-guiding rules (Allen et al, 2005 p. 149; Wallach & Allen, 2009 p. 84). A top-down ethical evaluation of an action or situation by a machine would consist in a reference to a higher-level rule or rules, and a judgement of the action or situation in light of this rule or rules.

Bottom-up alignment is essentially defined by Wallach and Allen as anything that isn't top-down alignment: *"…"bottom-up" approaches* [are those in] *which system design is not explicitly guided by any top-down ethical theory"* (2009 p. 101). This includes environmental processes for alignment, where we allow the machine to develop a motivational structure from acting in an environment we design. Artifical life modelling is such an instance, for which ethical behaviour has been shown to emerge (2009 pp. 101-106; Sullins, 2005). They argue that it is best to attempt to use both approaches in the design of what they term *hybrid* machines (2009 pp. 117-118).

Wallach and Allen's distinction here is quite similar to Bostrom's. If we compare Wallach and Allen's wording in defining bottom-up approaches to Bostrom's in defining direct specification, they seem like natural opposites:

> "…*"bottom-up" approaches* [are those in] *which system design is not explicitly guided by any top-down ethical theory…*" (Wallach & Allen, 2009 p. 101).

> "*Motivation selection can involve explicitly formulating a goal or set of rules to be followed (direct specification)…*" (Bostrom, 2014 p. 169).

The precise extent to which these distinctions are comparable is an exegetical matter, and not my central concern here. But it should be emphasised that indirect normativity can be top-down, whether or not it is top-down in the precise way that Wallach and Allen mean by top-down alignment. Proposals like MR and CEV are broad motivational goals. Loosely phrased: DISCOVER AND IMPLEMENT THE MORALLY CORRECT VALUE SYSTEM and DISCOVER AND

IMPLEMENT THE COHERENT EXTRAPOLATED VOLITION OF HUMANITY. These goals are not immutable, and they or may not be explicit. The AGI will be engaging in adjustment and potential revision of them, and it will *certainly* be engaging in the adjustment of goals on the ethical level of consideration. But the central attempt in these cases is an instance of top-down system design.

With value alignment it can be tempting to revert to a level of direct control over certain aspects of this project, a level which is appropriate for non-agential technologies. But with AI, and in particular with AGI, the functionality of the system is partly out of our hands. In narrow soft computing methods we don't explicitly tell the system which observations and associations to look out for. Instead we allow it to define and attend to features of its informational landscape of its own choosing, and to structure itself in such a way as to take full advantage of these features. Indirect normativity is a special case of this functional offloading where the objects of consideration are the values of the system itself.

Chapter conclusion

The development of AGI is unprecedented in that we are attempting to create a domain-general cognising agent. I have examined the potential of such an agent to pose a significant risk to humanity, in virtue of the capabilities and inclinations which we can reasonably expect it to have. Once an AGI reaches a critical point where it is capable of self-improvement, there is a significant possibility of an explosion of iterative self-improvement, resulting in a superintelligent agent. In the service of its prime goals we can expect, unless we intervene in the right way, such an agent to converge on several instrumental goals. These include, but are not limited to: self-improvement, self-preservation, resource acquisition, and the preservation of supergoals themselves.

Ideally any AGIs we build would not engage in efficient and effective optimisation of some non-humanitarian goal at the expense of humans. AGI alignment, the attempt to build an AGI whose goals and values are aligned with our own, allows for several strategies. Many of them fall into one of two camps. Direct specification involves the strategies that attempt to formulate motivational and axiological frameworks that will form the core value system of the AGI. Indirect normativity involves the strategies that offload the task of formulating such a system to the AGI itself by implementing the inclination to learn the appropriate values rather than the values themselves.

Existing critiques of direct specification point out how unlikely it is that we can codify ethics to a satisfactory degree, given our historical inability to reach widely accepted conclusions regarding the conceptual structure of ethics. These concerns are significant, particularly when we consider how inadequate ethical beliefs were even in the recent past. In the following chapter I will make a different argument, which supports the *in principle* unviability of direct specification. The issues I will highlight should be taken into consideration against a backdrop of potential superintelligence, where subtle issues have the serious potential to be magnified.

II: Points of failure

My aim in this chapter is to construct an account of the failure points in direct specification, and their inevitability. Earlier we considered Bostrom's distinction between final and instrumental goals. Here I would like to introduce a similar distinction, namely that between final and instrumental values. Values are evaluative concepts (Tappolet & Rossi, 2016 p. 4) which drive behaviour (Jiga-Boy et al, 2016 pp. 244-245). The value a person places on money can be sufficient to cause them to enter the workforce, and the value a reinforcement learning algorithm places on "reward" causes it to maximise its reward function. An instrumental value is the kind of thing that is valuable as a means to some further end, and a final value is the kind of thing that is the end itself (Korsgaard, 1983).[12] In other words, final values are those which need no reference to further values in order to ground themselves axiologically. Insofar as any agent is goal-directed, it can be understood as having values; evaluative concepts which determine which states of affairs[13] are preferable, and which drive actions which attain preferable states of affairs. I will ultimately claim two things: that direct specification can, at best, align the AGI's final values with our instrumental values, and that it is problematic for this reason.

Even outside the bounds of AGI alignment, people are playing the same game when they look for the best means to shared axiological ends. Where the ends themselves are distinct, they are not playing the same game. To illustrate: some political differences reflect

[12] Korsgaard distinguishes between a kind of agent-centric conception of value, and an object-centric conception. I am using the agent-centric conception; i.e the way something is valued: in its own right, or as a means to an end. The object-centric conception is more about the actual conditions for which something is valuable: intrinsically, or in relation to other things.
[13] I am using "states of affairs" as a general term; it can refer to world-states external to an agent, or states of the agent itself (like virtues), or actions, and so on. The precise things that evaluative concepts determine intrinsically preferable will depend on the particular instantiation of values.

differences in final values, and some reflect differences in instrumental values. Take the example of Eleni and Jocelyn, who both acknowledge that global aid is beneficial to the recipient nation. Jocelyn does not support aid on the basis that she wants a government to help its own citizens first and foremost, while Eleni supports it because she wants governments to help people regardless of citizenship, or to counteract historical factors that gave rise to the need for aid. Now take Amanda and Gerald, who hypothetically share the same core values of helping others. They disagree on the issue of raising the minimum wage. Amanda believes it will do more harm than good because of higher unemployment and higher prices, while Gerald believes that it will do more good than harm because it distributes resources more fairly between members of the economy, which has positive effects for the poorest members.

There are, of course, many more factors at play in these disagreements, but there is something to be said for the instrumental differences of opinion between Amanda and Gerald. There is a meaningful sense in which Amanda and Gerald are playing the same game, where disputes can potentially be resolved in a space where each is receptive to epistemic growth. Eleni and Jocelyn can agree on all the non-moral facts involved but retain their dispute. Amanda and Gerald are *aligned* in a way that Eleni and Jocelyn are not. AGI alignment, I claim, should be oriented towards alignment of this nature: an alignment of final values rather than of instrumental values. Later we will examine why this is the case, but first we must identify the failure points in the conceptual structure of direct specification.

Philosophical failure

Yudkowsky distinguishes between *technical failure* and *philosophical failure* (2008 pp. 318-319). Technical failure is the failure to build an AGI that works the way we want it to work. There could be some error in the architecture of the machine mind, or a mathematical oversight when designing a utility function, or any number of such issues. Philosophical failure is a failure in determining what kind of AGI we should build in the first place. For instance, say I successfully built Bostrom's paperclip maximiser. I may have succeeded *technically*, but in Yudkowsky's sense I have failed *philosophically* in building an AGI with such a foolish goal. The distinction could be loosely described as a failure to build what we want vs a failure in what we want in the first place, though this loose formulation conceals the source of philosophical failure, which is rooted in instrumental, not final, values.

Yudkowsky uses the example of political ideology to illustrate philosophical failure. It could be that one programmer is convinced that utopia would result from the implementation of

some version of communism, and another is convinced that utopia would result from the implementation of capitalist libertarianism, and so on. The successful implementation of whatever political system a given programmer prefers clearly has the strong potential to be problematic. Yudkowsky points out that these beliefs are the combination of value judgements (including what it means to be in a utopian state: everyone being as feasibly happy as they can be all the time, or everyone being able to live in the most socially uncoerced way possible, and so on) and empirical beliefs (including beliefs about the best sociopolitical structures that will bring about such a utopia) (p.320).

At the very least, we should attempt to program AGI goals that can account for changes in *empirical* beliefs. Goals which implicitly combine empirical beliefs and value judgements will prevent the AGI from properly adapting its behaviour with more information and intelligence; the philosophical errors of the programmers will retain even across empirical paradigm shifts of the AGI in question. An AGI that fundamentally wants to implement a particular political system will be disinclined to implement a different political system even if good empirical reasons were to emerge in its favour. An AGI that fundamentally wants to implement a utopia would be inclined to alter their instrumental political goal in light of new evidence. Of course, the question of what utopia means is a problem too, and this question is not, or at least not entirely, empirical.

Frames and points of failure

I would extend Yudkowsky's self-admittedly informal and fuzzy categorisation a step further. In addition to the philosophical failure of the instrumental fusion of final values with empirical beliefs about the world, there are *conceptual* points of philosophical failure, which can be understood as problems translating between frames. These conceptual points of failure are the terrain on which direct specification collapses. To understand why we must take a step back and examine the concept of frames from a conceptual and cognitive perspective.

The term "frame", as I use it here, describes a way of thinking about how both we and AGI represent things. It refers to a data structure, particularly a cognitive data structure, comprising a hierarchy of nodes (Minsky, 1974[14]; Anderson et al, 2006 p.42; Coulson, 2001 pp. 19-20). Minsky describes the idea: "*A frame is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's*

[14] From a memo accessed online at http://web.media.mit.edu/~minsky/papers/Frames/frames.html. Reprinted in *The Psychology of Computer Vision*, P. Winston (Ed.), McGraw-Hill, 1975.

birthday party… We can think of a frame as a network of nodes and relations. The "top levels" of a frame are fixed[15]*, and represent things that are always true about the supposed situation. The lower levels have many terminals–"slots" that must be filled by specific instances or data. Each terminal can specify conditions its assignments must meet. (The assignments themselves are usually smaller "sub-frames.")*" (1974).

This last bracket introduces the *recursiveness* of frames (Anderson et al, pp. 44-45; Peterson, 2015 pp. 44-45; Barsalou, 1992 pp. 35-37). In essence, sub-elements of frames are also frames, as are collections of frames which all comprise a constitutive hierarchy. This is one way to understand knowledge representation in cognition and neural networks (Martins et al, 2017; Russell & Norvig, 2016 pp. 437-442). Constitutive hierarchies are those in which super-elements are comprised of sub-elements; think *humans* and *cells* or *celebrations* and *birthdays*. For the latter case we can think of celebrations as a set of events which is comprised of birthdays, graduations, weddings, and so on.

In cognitive terms we can cash out this model of hierarchical knowledge representation in terms of levels, which depending on the grain of analysis can be instantiated in *networks* (Supulchre et al, 2012; Bastos et al, 2012) or *layers*, which are comprised of individual nodes which have a higher incidence of interconnections among them (Dumoulin, 2017; Bishop, 2006 pp. 225-269). This cognitive hierarchy is not a constitutive hierarchy. There is, however, a conceptual overlap between the cognitive hierarchy, i.e the organisation of networks and layers, and the constitutive hierarchy of knowledge representation. Higher levels of the cognitive hierarchy are responsible for the broader, more widely-applicable concepts that are represented in the brain, while lower levels of the cognitive hierarchy are responsible for the narrower, more situational concepts which comprise the lower levels of the part-whole hierarchy (Martins et al, 2017).[16] The significance of the distinction and overlap between these two hierarchies will become apparent.

Frames are a description of the informational *contents* of this neural organisation. In the context of ethics, and particularly of the project of AGI alignment, we can divide spaces of ethical theories, reasoning, and expression into various frames. Hill Jr's (2000; 2012) conception of high-level and intermediate-level moral considerations illustrates one dimension of ethical framing. For a Kantian, concepts such as duty, the universal law formula, and the formula of humanity are high-level concepts. They can be used to judge the

[15] It should be emphasised that they are only fixed in terms of the frame itself: if they change we have a different frame. In viable systems that employ frames, unless those frames are explicitly and immutably specified, they will be subject to change as the system learns.

[16] The part-whole hierarchy includes certain specific accounts of the kind of abstraction that occurs, such as transformational invariance (Buckner, 2018) and temporal (Kiebel et al, 2009), because representation of longer time-scales (and invariance) is comprised of the invariant statistical features of fast-changing stimuli.

moral status of intermediate-level concepts like lying, justice, and friendship. For the rule utilitarian, the high-level concept is the principle of utility which is used to determine which intermediate level rules should be adopted. Decisions in particular circumstances are low-level, and are made with reference to the intermediate level rules (Hill, 2000 p.12-13).

Just as the goal "be good" requires consideration of empirical information pertaining to mental states, communication, cultural dynamics and so on, that same goal requires consideration of lower level theoretical framings. "Thin" moral concepts like "good" and "wrong" bear certain relations to "thick" concepts like "kind" and "just" and "painful" and "well-intentioned", though the nature of these relations is in question (Elstein & Hurka, 2009). Thin concepts refer to vaguer, more widely applicable concepts while thick concepts refer to more descriptive and therefore narrower concepts. Thick and thin concepts are on different levels of consideration. In carving up the moral landscape in coarser or finer grains, they are different frames.

The reframing of a high-level descriptive concept (a thin concept) to a low-level descriptive concept (a thick concept) is a recursive step in terms of the constitutive representational hierarchy. Crucially, however, it is a *non-recursive* step in terms of the cognitive substrate of the frames. This is the root of conceptual points of failure. To illustrate: I could say that the concept "autonomy" is constituted by several thicker concepts including "consent" and "freedom of movement". Cognitively speaking, my concepts of "autonomy", "consent", and "freedom of movement" are all represented in different places, or with different functional patterns of activation. So the cognitive move from "autonomy" to "consent", and consequently the semantic move, is a move between frames (or sub-frames, depending on the level of consideration). This poses a serious problem for direct specification; I will examine why in more detail shortly.

Instrumental values axiologically constitute final values

Descriptively, the highest-level moral concepts are the thinnest, the most abstracted from lower level concepts, and the largest structures in conceptual spaces. *Axiologically*, the highest-level concepts are final values, and lower level concepts are instrumental values; we determine which instrumental values to adopt with reference to their conduciveness towards final values. While these senses of hierarchy are certainly distinct, they overlap to an extent in local, conceptual considerations of moral principles. If we take a final value to be the central abstraction in an analysis, the concepts that are instrumental (i.e lower level) to the

final value will relate to the final value in that the final value abstracts over them in an axiological sense. For instance, if we want to say that justice, happiness, and honesty are each good, then good is a broader category of which justice, happiness, and honesty are members. If, hypothetically, our final value is goodness, then justice, happiness, and honesty are lower-level *instrumental* values; there could potentially be a conceptual (or empirical) argument that disconnects any of them from the broader concept "goodness". And we adopt them as values under the condition that they are good.

But this example is too easy: in this case justice, happiness, and honesty happen to conceptually constitute goodness, among other thick concepts. It could be that the thicker concepts of justice, happiness, and honesty are our final values in and of themselves, in which case goodness becomes a mere abstraction from them. But even in this case, *along the axiological dimension* the final values are still the constitutively higher-level frame. If I have such final values and am reflecting on why I (instrumentally) value goodness, I will refer to the overarching final value frame and will see that goodness is valuable *in that* it is an abstraction of the final values. In other words, while the final values in this case (justice, happiness, and honesty) conceptually constitute the instrumental value (goodness), the instrumental value axiologically constitutes, in part,[17] the final values. The final value frame consists of concepts which themselves have recursive associations with other concepts; to instrumentally value goodness is to refer to the overarching final values that goodness is associated with. So instrumental values are lower level on the axiological dimension of description, even when they are higher level descriptively.

Let us consider another example. Say that, hypothetically, my only final political value is the happiness of individual citizens. Political systems exist on a constitutively broader level of *descriptive* analysis than the happiness of individuals. But in this case, my axiological system of frames will have as its highest element the happiness of citizens; the value I place on different political systems is subsumed under this consideration, with each system being subject to this overarching factor. This is, naturally, a consequence of framing. While we can describe a constitutive relationship in one direction under a particular dimension, such as descriptive levels of analysis (i.e citizens and political systems) or conceptual constitution (i.e goodness and justice), under a different dimension this constitutive relationship can reverse or be deflated.

[17] When I say that x, y, and z constitute w, I do not mean that they exhaust the constitutive fabric of w. For instance, there is more to "goodness" than justice, happiness, and honesty; I am merely illustrating the conceptual part/whole relationship between them rather than making any claims about those concepts being *all* that goodness is comprised of on that level of description.

Frame mappings

In AGI alignment, and more broadly in the study of ethics itself, there are many potential points where mappings between frames must be explicitly or implicitly constructed. The construction of such mappings is one of the fundamental tasks of the ethicist. The points at which mappings must be constructed are potential points of failure. The implication here is that mappings can be unsuccessful, which itself can mean different things depending on the criteria of success, criteria which are themselves sensitive to the various metaethical perspectives one can hold.

A point of failure has the potential to bring about such outcomes in virtue of a failure to map from our final values to the actual AGI actions that may violate those values. This can occur on any of the many steps involved in implementing an aligned value system into an AGI. To illustrate: the process of building an AGI which does what I want will involve, in the first instance, a (rough or precise) notion of what I want. I would then need to reframe what I want in terms that comprise a normative framework, and then to reframe this in terms suitable for an AGI goal. This is then reframed as a mathematical expression which will be the ultimate utility function of the AGI. From this motivational structure, and within a certain context and with a certain level of cognitive capacity, the actions of the AGI will emerge, which will ideally align with what I want.

Conceptual failure in direct specification

What does direct specification involve? A precise answer depends on the particular kind of specification. A set of rules is a different beast to the implementation of, say, a hedonic utilitarian supergoal. But something they all share, I will argue, is that the goals which are specified are instrumentally valuable; a means to an axiological end. The axiological end in question is *what we actually want* out of AGI, what we hope to achieve, and what we hope to avoid.

When we decide to implement a set of rules or goals into an AGI which we hope will attain the kind of result we want, we are engaging in a means to an end. In terms of values, we can say in this case that the AGI has values which are instrumental to our final values, whether they are successful or unsuccessful at doing this. An overt case of human-instrumental AGI values would be, following Yudkowsky's (2018 p. 319) example, the implementation of the

goal to enact a particular political system. This goal conflates empirical facts of the matter with values, such that implementing a particular political system can only be appropriate if we happen to be right about the empirical facts. If we successfully align the AGI with our final values, we will expect it to implement[18] this political system anyway if we were empirically correct, and to implement a better system if we weren't.

A similar problem can arise in the conceptual sense. Just as a relationship between *x* and *y* can be instrumental in that *x* causally leads to *y*, it can be instrumental in that *x* conceptually leads to *y*. An example of the latter might be an attitude like: "if I act with honesty, courage, and selflessness that *means* that I am a good person". Assuming that being a good person is my final value, conceptual links and breaks can be drawn from the lower-level concepts; for instance, one might argue that selflessness is not a quality of a good person, and that a healthy consideration for oneself as well as for others is key. This is only possible because of the conceptual distance between "good person" and "selflessness".

Instrumental values are those which are valuable in that they attain final values. This applies to the conceptual domain, as we have examined: some concepts are descriptively constitutive of other concepts, and some concepts are axiologically constitutive of other concepts. Translation into entirely new frames subverts the concept in the *descriptive* sense; we would say "x *means* y" in much the same way as we might say that "autonomy *means* respect, freedom, agency", or some appropriate definition. The disconnect between the original concept and the translation is precisely what produces the means-ends dynamic. This gap is more descriptive rather than *directly* axiological, but the descriptive gap belies an axiological gap: the value of concepts supervenes on the concepts themselves.

Aligning the AGI with final values rather than instrumental values is the better option unless we are confident that we will be better at determining the facts of the matter, whether empirical or conceptual,[19] which is *especially* unlikely at the point where an AGI has the capacity to be doing things like implementing political systems, and so on.

Of course, direct specification can involve the attempt to directly codify our final values, and load this into an AGI. I will call this *final value specification*. However, in so doing it remains an implementation of instrumental values. This is the case because the codification of final values is itself a form of instrumentation. Insofar as codification is a means to an end, being held to a standard of validity external to itself—how accurately the codification captures our final values—it is a form of instrumentation: it merges what we want with conceptual facts about how we attain what we want. In a similar vein, the *results* of the codification are

[18] That is if implementing a political system at all is an action aligned with our final values.
[19] Conceptual facts of the matter would regard coherence and consistency of local systems, as well as the accuracy of translating between frames.

instrumental values. The result of the codification will necessarily be a different frame, existing with a different conceptual structure, instantiated in a different cognitive structure. These results are likewise accountable to the external frame which are the original final values that we have attempted to codify.

To have a final value, and to attempt to express that final value in a different semiotic system, involves *at least* one conceptual degree of separation. Consider what it would take to design an appropriate utility function for an AGI, even if we knew what we wanted it to value from the first instance. The kinds of things that our final values are would need to be expressed in terms of some mathematical formalism, or an otherwise specification-appropriate framing. This is not necessarily a fundamental obstacle when discussing certain formalisms with other humans. We can point to some variable and say "this represents the net utility that will result from an action" and "that represents the preferences of an agent at time *t*". The framing issues are concealed by a relatively clean reference to the semiotic system we already work with, which has a wide range of assumptions underlying it: what we understand utility to be (happiness, pleasure, preference, etc), the things we understand as agents, and so on. For an AGI, none of these assumptions will be present at the outset; the problem of defining the vast semantic web of each moral concept we use is still ahead of us. That is unless we engage in indirect normativity, and any tentative value specifications are understood as attempts to capture a dynamic which is itself the true target.

Final value specification involves at least one, though more likely two key steps of abstraction away from final values.[20] Each step of abstraction is a potential point of conceptual failure. The first step is from final values to a natural language expression of final values. The second step is from the natural language expression of final values to a form suitable for an AGI goal, likely a mathematical formalism.

The first step: expressing final values in natural language

This first step is a requirement for attempts at final value specification: we need to explicitly understand, and communicate, the values we are trying to capture. To understand this step, we must first consider what human final values are, in the sense we might need in order to construct a system with our final values as a reference and an aim. My aim here is to be as neutral as possible with respect to existing and potential research on which values are

[20] One of the steps may not be necessary, which I will discuss shortly.

pragmatically best to use, whether directly or indirectly, to inform AGI goals. For instance, proposals that suggest a form of Inverse Reinforcement Learning (Hadfield-Menell et al, 2016; Hadfield-Menell et al, 2017; Mindermann et al, 2018) understand final values as the reward function of an agent, a function which can be inferred through observation of the agent's behaviour. Proposals like Yudkowsky's Coherent Extrapolated Volition (2004) aim for a collective ideal observer account of human values. Neutrality with respect to various conceptions of values is important when mounting a critique against something as broad as direct specification; all that matters for the critique is the non-equivalence of our starting point, the values which we aim to transfer to the AGI, and the actual values of the AGI.

I will begin with a tautology: only the final values which are identical to linguistic expressions are identical to linguistic expressions. Any conception of final value as anything other than linguistic necessitates a conceptual removal from the linguistic, which itself necessitates the requirement for a mapping between the linguistic and the final value. In other words, only explicitly linguistic final values can avoid the need to bridge the gap between expression of value and value.

To clarify what is meant by "linguistic final value": final values, in the way I understand them, are mental concepts, and if a person's mental concepts are words that they have adopted as representations of moral principles or world states, then the move from value to expression has been made from the outset. Say a person grew up in a familial and cultural environment where everybody strongly valued honesty, and they internalised honesty as their final value. Their understanding of honesty would be drawn from the word as used in their linguistic context, from the patterns in the particular ethical cases described as instances of "honesty" or "dishonesty", and so on. These experiences would build and shape their mental representation of honesty in a way that is aligned with the external cultural understanding of honesty, and more precisely aligned with their own expressions of the term "honesty". For such a person, talk of the final value of honesty in natural language likely does not entail much of a conceptual gap between expression and reality of the final value. This view has some neurological basis (Zahn et al, 2009; Feldman Barrett, 2018); linguistic conceptualisations of evaluative attitudes have a neurological consistency even across varying situations. This does not imply that the concept-words that are robustly associated with evaluative attitudes are *final* values, however; they could be instrumental values. Coherence accounts of final values in the sense of the goal of alignment, like CEV, are likely to map onto this linguistic conception of final values even more cleanly, as words can be understood as conceptual coherence between members of a linguistic community. If the linguistic conception holds, however, there is still another step of abstraction that is necessarily involved in direct specification, which I will examine shortly.

Non-linguistic conceptions of final values would entail a gap between expression and reality of the final values. The expression of final values requires words, or other signs, which we take as referring to the final values. The precise nature of this reference, this bridging of the gap, will depend on the nature of the expression, and the nature of the final values. There are two potential issues with the bridging: where the expression does not capture the full nuances of the final value, and where the expression implies additional features that are not part of the final value.

Any expression of a set of final values must be made within some frame, where the relevant concepts are defined with respect to other concepts. The Wittgensteinian approach to meaning would, in an ethical context, take the concept to be defined in terms of its usage, which is to say in terms of the particular cases where we apply it (Wittgenstein, 1969 p. 20; see also Grim et al, 2004). If our final values are on a more abstract level of specification than particulars, the expression of them in some natural or other language is entangled with the particulars of world states and actions, which are themselves taken to be instrumentally valuable. Thus it may not be possible to extricate means from ends for such expressions. This is a problem if we are trying to give the AGI as much space as possible to revise epistemic beliefs. To a 19th century doctor, helping a patient may have involved prescribing them heroin. Were such a particular to be included as part of a definition of "helping", it seems that the conception of helping may be corrupted.

This step also poses a problem for our ability to grasp and express our own final values explicitly. An example of a conceptual failure here would be a person expressing that their final value is to make everyone as happy as possible. If they would also balk at the idea of consigning everyone to experience machines (Nozick, 1974), hypothetical machines which constantly simulate experiences which are maximally pleasurable for those who are plugged in, then they have failed to translate their actual final values into natural language. This step, though a requirement for any ethical discourse, is incredibly difficult to get right. But to engage in final value specification, final values must be consciously understood, which entails their expression in natural language.

The second step: from natural language to AGI-appropriate goal

The form in which a goal must be expressed to be suitable for an AGI is not in natural language. It is important to distinguish here between "loading" a goal into an AGI in the process of development and giving a goal to an AGI that already exists. In the latter case, if

the AGI is inclined to obey, and capable of understanding natural language, then natural language goals are fine. This latter case, however, requires an AGI that is developed enough to be able to understand natural language goals, and which is inclined to obey our instructions—we would have already solved the alignment problem in this case. Natural language instructions would then be instrumental goals for the agent. The process of creating such an aligned AGI is where my current concern lies, and to get there we must reframe appropriate values in a form appropriate to the AGI engineering process.

This form will likely be a mathematical formalism, or perhaps an element of some new symbolic system developed for the purpose of AGI cognition. I will simply refer to AGI-appropriate forms of specification as "formalised". The conceptual distance of the gap between what we attempt to formalise and the formalisation itself will depend on the former; certain starting points entail a greater gap. If we begin from ethical concepts and frameworks, the gap will be significant; ethical concepts are instantiated in a web of meaning, with semantic and referential properties that are distinct from the space of formalisms.

Formalisms, insofar as they are mathematical and/or logical expressions, are a priori structures that relate mathematical objects and properties between one another. Human evaluative concepts are instantiated in cognitive frames where reference to external objects, properties, and abstractions over those objects and properties have been shaped by years of association with the world and with conceptual frames. The gap between a formalism that we might construct, and the evaluative concepts which inform such a construction, is going to be significantly different in structure, unless the mapping is made in great, and explicit, detail. This is a task better left to the AGI itself, as I will discuss shortly.

The problem of reference is even deeper. When we communicate our evaluative concepts we are drawing on a web of implication, of semantic grounding, and of reference. To say "the ultimate good is the maximisation of the pleasure of agents" requires an implicit understanding of the meaning of these words, and of the (potentially quite distinct) meaning of the sentence. To be fair, arguably some of these meanings do translate relatively cleanly to formalisms. "Ultimate" and "maximisation" can be framed as *amounts*; the amount of utility, and the amount of "preference". But "pleasure" itself, as well as "agents", may be significantly more difficult. Some proposals, as discussed earlier, reframe ethical concepts such as "preference" as the utility function of an agent. This may be our best bet, though it comes at a cost of determining the best ethical instantiation on the basis of what is easy to formalise. However, there is *always* a distinction between the formalisation and our actual evaluative concepts, even if we can find mathematical analogies from some concepts more easily than others.

The deep structural and referential issues in both the steps we have examined necessitate that direct specification will result in an AGI that has final values which are not equivalent to our own, even when we attempt to directly specify our own final values. Insofar as the translation of final values is itself a means to an end in that it is held *accountable* to the external standard, the end result is a system of our instrumental values. Recall the above discussion about frames, particularly the differences in cognitive instantiation. When a final value is codified in a different cognitive and semiotic frame, this concept loses the equivalency, even if there is a clean constitutive relationship between the preceding and resultant evaluative concept in the isolated ethical sphere. A gap is created which entails an instrumental relationship; a means-end relationship to the original concept. A conceptual distance is always present in the reframing of final values to new semiotic spaces. Even if the direct specification can avoid fusing *empirical* means with our ethical ends, it can't avoid fusing conceptual means with axiological ends.

Underspecification of final values

Practically speaking it may not be possible to succeed in the project of directly specifying an expression of human final values, if our final values are underspecified. In this context, "underspecification" refers to a high-level goal which has multiple consistent low-level interpretations. One example is an implicit evolutionary goal such as "survive and reproduce", which can be achieved in many ways. I do not mean to suggest here that "survive and reproduce" is a human final value; it is simply an example of an underspecified high-level goal. For humans, low-level behavioural coherence could be established by virtue of evolved or otherwise distinctly human cognitive tendencies that lend towards a certain family of lower-level interpretations of the high-level ends. In humans this particular evolutionary "goal" has likely resulted in pro-social moral instincts which help attain it, though this is not a necessary means to such a goal; there are other ways of surviving and reproducing.

In such a case, our final values would be combined with a faculty of *judgement*, which determines the application of general principles (whether final values or instrumental values) to particular instances. A literal thinker may take an ethical framework and algorithmically apply it to extract contextual action-guiding principles, but judgement is what allows them to do it *in the right way*, which is not necessarily able to be defined through isolated consideration of the framework as the framework. Aristotle takes the golden mean to be determined "*by a rational principle, and by that principle by which the man of*

practical wisdom would determine it" (EN 1107a1-3). On this account the principle alone is insufficient.

Thiele draws a distinction between top-down algorithmic applications of generalities and judgement, which is a kind of moral common sense (2010, 95-104); an intuitive, bottom-up, affective evaluation of a situation. In terms of AGIs, this could be a cognitive architecture that is designed to be suited to the evaluation of moral considerations, with reference to the final goal. Even if there are multiple possible conceptual structures that are consistent with a particular supergoal, and multiple possible standards of right action resulting from those conceptual structures, a particular cognitive structure could shunt the AGI down a certain path of implication. The problem of designing such an architecture, however, may be an AGI-complete problem. In other words, it is likely that we will need to understand how to build general intelligence before we can begin to implement any form of judgement based on cognition if judgement is rooted in the cognitive profile of the agent. AGI alignment should precede knowledge of how to build AGI, so we may need to look elsewhere for the time being.

What's so bad about instrumental values?

Even if the claims made above are correct, the proponent of direct specification could simply accept that it will necessitate an abstraction away from our final values. We are generally content with designing tools with functions that are instrumental, and even then, only instrumental for a subset of our preferences. The issues that I have identified are by no means isolated to AGI alignment. They apply whenever we attempt to translate, whenever we communicate with other people, whenever we give instructions, and whenever we design tools and task-specific algorithms.

These issues are such a concern in this context because of the uniquely high standards we must have for AGI alignment. The most significant risks of AGIs are those associated with the possibility of superintelligence. While individual humans, tools, and algorithms often engage in behaviour or functionality which I would consider unaligned with my own values, and the risks involved can potentially be significant, none are quite as significant as the risk a superintelligence could pose. Tools and narrow algorithms do not have the ability to engage in domain-general cognition and behaviour; their unaligned functionality does not generally run the risk of ballooning out of proportion with a superintelligent resilience to any attempts to subvert that functionality. There is no restricted mode of operation of an AGI; its cognitive

reach is theoretically boundless in terms of the spaces in which it can operate. For narrow algorithms, "goal" specification can ignore the vast majority of morally relevant details as they will simply never be a factor in their operation. We do not have that luxury with AGIs. Values which are not robust across a wide variety of spaces, including spaces that we have no current understanding of or experience with, can potentially lead to unaligned decisions in those spaces.

Take, for instance, the case of an AGI which has the prime goal of accurately answering our questions. Answers to the questions we ask are things we likely only value instrumentally, but we can work with them well enough. However, as Bostrom (2014, p. 150) notes, were we to ask a question which required a significant amount of computational resources to answer, an AGI would be motivated to convert as much matter as it could into *computronium* - a hypothetical arrangement of matter in a form optimised for computing. This conversion of matter is a route to goal completion that is accessible to it in virtue of it being a general problem-solver. Naturally we could add caveats to our specification like "do not convert matter into computronium", but we are limited to patching the holes that we can identify with hypothetical examples. These holes are not limited to non-ethical goals. The history of ethics is rife with thought experiments and hypotheticals which purport to show that a consistent application of a particular ethical theory would result in something that even proponents of that theory would disagree with. Following this is a subsequent patching-up of definitions to exclude or otherwise evade these eventualities.

Particular to instrumental value specification is the gap between what we want and what the AGI wants. This gap is a locus of insensitivity to human final values. Ideally we should like an AGI to develop new action policies, in light of epistemic growth, that better attain our final values. Ideally we should not bake our own fallible epistemic stances into the AGI values, because this will make the AGI resistant to the right kind of growth, the kind that we should prefer as it allows us to attain what we want. A final value, perhaps a utility function or some other instantiation of a motivationally fundamental set of rules, goals, and constraints, is intrinsically motivating to the agent in question. These final values do not reduce to other motivations, whereas instrumental values are only valuable in their capacity to attain final values. The concern here regarding AGI alignment is how we should think about the final AGI values as connected (or identical to) our own values. Here we are interested in the human axiological case. Ideally we should like to align AGI values with ours in terms of our final values, and allow it to develop and reflect on the best means to those ends.

Does indirect normativity avoid these problems?

No, it does not. But there is a crucial difference that has the potential to vastly improve the robustness of indirect value selection methods compared to direct specification. In virtue of the definition of indirect normativity as allowing the AGI to learn the right values, the gaps between frames that poses such an issue to specification will not be immutable. An AGI will learn how to narrow those gaps over time, as new conceptual and empirical knowledge is gained.

The objection could be raised here that direct specification is not necessarily immutable. One way this could potentially be achieved is through implementing an AGI that has an explicitly specified set of values which include the inclination to allow us to revise its own values. This proposal may seem reasonable, but there are some difficulties to consider. To begin with, the revisable values must not be the AGI's final values. Any final value that an agent holds will incline the agent towards preserving the final value (Omohundro, 2008). Any potential revision to the final values of an agent in state S will be considered with respect to the agent's values within that state. An agent's potential values at state P may be better fulfilled with the potential revision, but it is agent S that is considering whether or not such a revision is appropriate, which must be according to S-values.

There may still be some resistance to this argument in the form of a revision that appeals to S-values. Consider the case of an AGI who, for some reason, highly values stamps.[21] Suppose that we were to propose an alteration to this AGI: in exchange for a million stamps, we will give it the inclination to also value pens, to a mild degree. Furthermore, the AGI trusts that we will alter it exactly as we say we will. To the AGI at state S this might seem like a good deal. Say that they accept this offer. Then, with its new values in place, we give it a new offer, to again increase its valuation of pens slightly for a million stamps plus a million pens. The revised AGI, particularly with its newfound mild taste for pens, is amenable to this arrangement. And so we keep iterating this process until the AGI highly values both stamps and pens. Following this, we repeat this process in the reverse direction – at each step offering the AGI some large number of pens so that we can slightly reduce its valuation of stamps, until eventually it doesn't value stamps at all.

At each step of this process, the isolated consideration of the offer seems to conform with the AGI's values. Knowing the entire process, or at least the possibility of such a process, the original AGI will be inclined to reject the offer entirely. This stems from the convergent

[21] Inspired by the parable of Murder-Ghandi from https://www.lesswrong.com/posts/SdkAesHBt4tsivEKe/gandhi-murder-pills-and-mental-illness

instrumental goal of supergoal preservation (Omohundro, 2008 pp. 5-6). To counteract this, we would need to design the AGI such that it is inclined towards value adjustment, and value learning. If we succeed here, then value revision proposals like those mentioned above are possible. But even so, it would be better if we offloaded such revisions to the AGI itself.

Offloading translation

Bostrom originally motivates indirect normativity with reference to his "principle of epistemic deference" (2014, p. 258). This principle essentially states that, given the greater cognitive capabilities of a superintelligence, it is more likely to be correct than we are on a range of issues, and so we should defer to its judgement when possible. This seems reasonable enough for most issues, though I would retain the human pre-eminence when determining the axiological ends we are aiming for.

I would extend this principle of deference to AGI in general, not just superintelligence, in the particular domain that concerns the conceptual gaps between frames which I have been thus far pointing to as a core issue in value alignment. This is because soft computing methods are ideal for constructing *arbitrary mappings* between frames. Arbitrary mappings are those where the relevant features and associations of the consequent frame are not implied by the antecedent frame. I will expand on this further after some preliminary remarks regarding translation between frames.

Fundamentally, moving between frames of consideration is an act of translation. A given frame has its own conceptual elements which relate to other elements of that frame in certain ways defined by the frame itself. The most abstract space of normativity involves concepts like "ought", "should", "good", "bad", "right", "wrong", and so on. A specific ethical theory will make use of these concepts in a sense that is designated by the theory itself. A hedonic act utilitarian will assert that "good" is determined by pleasure, and that "ought" in the sense of "I ought to do x" is determined with reference to the good, in some capacity determined by the principle of utility—namely maximisation (Frey, 2017). A Kantian deontologist will assert that "good" is a property of the will oriented towards duty (Kant, 1785; 4:393)[22], and that "ought" should be defined in the sense of the imperative to act in a way such that they could, without contradiction, will everyone to act in that way (Ibid, 4:421; Timmons, 2017).

[22] Here I use "good" to refer to a final good, not an instrumental good. "*Therefore this [good will] need not, indeed, be the only and entire good, but it must yet be the highest good, and the condition of everything else…*" - Groundwork 4:396.

Now consider the move from ethical theory to action. A particular action in a particular context can be described in ethical terms in the sense of how it comports with the relevant ethical framework. One action may be good to the hedonic utilitarian and bad to the Kantian; the action is being translated in different ways. The set of potential actions is made into a particular framework by the hedonic utilitarian, where they are evaluated with respect to their being able to bring about a particular world-state, where the ethical logic of various actions consists in their ultimate connection to pleasure, whereby they are weighed against each other in that respect. The understanding of moral actions to the Kantian is broader, where actions are weighed with respect to the motivations of the actor, and their comportment with the Categorical Imperative, and their status as permissible or obligatory.

These are moves between frames. Actions exist as actions regardless of which, if any, ethical framework is being applied. But the normative concept "ought" is a way to reach from the ethical framework to the action itself, and "ought" is contingent on the ethical framework itself. The space of actions is reframed in light of the ethical framework. Identifying mistranslations between the two are, in theory, a good way of getting people to act or think differently. Imagine two hypothetical charities: The Ineffective Institute and the Effective Foundation. One utilitarian may say "I ought to donate to the Ineffective Institute", and another utilitarian may point out that only a negligible percentage of their donation will go towards legitimate charity work, and that there is evidence of embezzling in the organisation. Maybe, they suggest, a donation to the Effective Foundation would lead to a more meaningful impact. This suggestion has the potential to be convincing in light of the misalignment between "I ought to donate to the Ineffective Institute" and "I ought to attempt to bring about the world state that is maximally good". There are many such examples of misalignment. An overt case would be a person who believes they ought to act in a certain way, and then proceeds to act in the opposite way. While their action was caused by a complex interaction of beliefs and associations, it is at any rate inconsistent with the ethical framework that they purport to hold.

Some mappings are almost inevitable in virtue of the antecedent frame's overlap with the consequent frame. For instance, the principle "never lie" can perhaps be said to robustly translate to "do not lie to Helen", because they share the concept "lie". This is because the particular case that is described as "lying to Helen" is understood explicitly in terms of its instantiation of the broader category of "lying". This too is a question of framing: if, rather than describing (and thus, perhaps, prescribing) the particular in terms of the general concept, it is instead described as something like: "telling Helen I didn't eat the biscuit", the connection to the frame that includes "lying" as a morally pertinent concept or category of actions must be somehow provided.

Some mappings are more arbitrary in that no connection seems forthcoming between two frames; the connections must be made "manually", as it were. Arbitrary mapping is the case for semantic categories in general - there is nothing inherent to a label in and of itself that necessitates a particular extension.[23] If moral principles operate in a similar manner - i.e if "injustice" is simply the semantic category that a range of particulars like [a policeman targeting, without having received a tip or other information, a random individual for a drug search on the basis of their race] and [a company refusing to reimburse a client for the company's error] and so on, then we are resigned to make all the connections between those frames ourselves.

Bottom-up alignment techniques can be of help in this matter: the association between a particular and a principle can simply be explicitly made without the particular being implied by the principle. For instance, we could engage in a kind of supervised learning where a training set of particulars is each assigned one or more relevant principles (possibly even as high-level as simply "right" or "wrong"). In general, the associative complexity, and sensitivity to information, of soft computing approaches makes them well-suited to construct arbitrary mappings between complex frames.

Codification and particularism

Thus far I have been arguing from a model which takes many values as being *reducible* to one or more final values, at least in an axiological sense. To understand why someone would instrumentally value something, we look at the ways in which it is an instantiation of, or a causal factor in achieving, their final values. From there we can model normative judgement in a hierarchical fashion. This axiologically reductionist model could be challenged, particularly by the moral particularist. In the words of Margaret Little: "*A deeply influential theme in the last few centuries of moral philosophy holds that ethical inquiry is the search for the architecture of morality*" (2000 p. 278). Particularists are those who claim that no such architecture exists, or that if it does it is just as complex, expansive, and holistic as the landscape of morality itself, thus serving as no epistemic convenience (Dancy, 1993 p. 56; 2004 p. 7, pp. 111-112).

Axiological reductionism, as I will call the approach I have been employing so far, entails that moral concepts, including values, are abstractions which have a consistent *valence* in

[23] There may be elements inherent to the usage of the label, and other facets of it, but the label itself bears no necessary extension within it.

lower-level considerations (McNaughton & Rawling, 2000 pp. 257-258). This is to say that a reason for a certain evaluation (once we traverse down the chain of reasons to the final value(s) that grounds them) is always a reason for a similar evaluation, even in a different context. Take the concept "pleasure". If pleasure is a proper sub-concept of a final value, which seems reasonable, and especially if pleasure is the final value itself, then we might say that something being pleasurable is always in favour of that thing. This is not necessarily a sufficient reason: it might always weigh the scales on one side of a decision or judgement, even if it is sometimes outweighed by considerations on the other side. Dancy makes the observation that sometimes pleasure can be a reason *against* doing something (Dancy 1993, p. 61). If a killer takes pleasure in the act, the pleasure *itself* is something most are morally averse to, among other features of the act. To someone who shares this intuition, pleasure is not a universally positively-valenced concept. If this is the case for all evaluative concepts, then the axiologically reductionist model doesn't hold.

Two possibilities seem apparent in response, if there is validity to these claims. The first is to admit that we really do have final values, but *any* traversal down the conceptual hierarchy will lead to concepts which are hopelessly entangled with other values, such that any form of axiological reduction is futile. In this case we could never model an instrumental value as being an expression of a final value. This is a stronger version of the case I have been making regarding instrumental values. The second possibility is that we do not have final values to speak of; that such a model is simply an inaccurate description of ethical reasoning, and that we should look elsewhere for a standard of success in AGI alignment. In defence of this second possibility, Nagel argues: "*This great division between personal and impersonal, or between agent-centered and outcome-centered, or subjective and objective reasons, is so basic that it renders implausible any reductive unification of ethics—let alone of practical reasoning in general. The formal differences among these types of reasons correspond to deep differences in their sources.*" (1979 p. 133).

In defence of ethical abstraction

There is at least one sense in which we cannot simply model ethics on the lowest level of description, with the interactions therein. Cognition is fundamentally hierarchical, with a division of labour between different cognitive levels that govern different levels of description (Martins et al, 2017; Dumoulin, 2017; Supulchre et al, 2012; Buckner, 2018). The way that biological brains and deep neural networks achieve the level of flexibility that they are capable of is through abstraction—reformulating low level information into models that

describe the information in broader and broader ways, producing predictions and policies relevant to the level of analysis they are concerned with, allowing us to act in ways that are sensitive to our short- and long-term goals in a changing environment, with respect to local- and global-pattern recognition and construction.

Throughout this work I have made reference to GOFAI—Good Old-Fashioned AI—and soft computing approaches to artificial intelligence. I am a firm believer in the disadvantages of the GOFAI approach, except where it is instantiated alongside and within the flexibility of connectionist networks (Wermter & Sun, 1998). But even in a GOFAI system, abstractions are a necessary component. As of this moment we, to the best of my knowledge, have no ability to create an intelligence system, nor even a system which is a pale imitation of intelligence, without employing abstractions. This is true for the space of cognition in general, let alone the subsets of this space which we call ethical and/or evaluative reasoning.

Even if we were to engage in alignment that was purely directed at the lowest level of ethical specification—actions in particular situations—i.e purely bottom-up alignment, a truly intelligent agent like an AGI would be formulating more and more abstract ethical concepts that make sense of the slew of low level information. Even where the evaluative concepts are not aligned with our own, they are still concepts which carve up the evaluative space on a level more abstracted than the particulars. Jackson et al (2000) argue that, as long as we are carving up the evaluative space into *right* and *wrong* (or *good* and *bad*, or *preferable* and *not preferable*), as is anybody engaged in ethical or evaluative reasoning, then there are abstract structures in the moral landscape which hold. This is necessary for any project which takes seriously the cognitive preconditions of evaluative judgement.

Guarini (2010) empirically expanded on Jackson and colleagues' argument by developing simple recurrent networks (SRNs) to classify moral situations (in the form of sentences) into one of three categories: permissible, impermissible, and uncertain. He found that there were patterns of unit activation that were involved in certain moral classifications, which *generalised* to new sentences. In other words, patterns were identified that seemed to reliably cause the same judgement in different cases; these represent evaluative abstracts, albeit simple ones. It could be argued that such networks are too simple to be directly analogous to human or AGI cognition, which is true. But it does show that the complexity of a connectionist system does not entail a similar complexity in the relations between abstractions and particulars; this in turn entails that modelling such relations adequately is not going to require a map which is just as detailed as the territory, as Dancy (2004) has claimed.

Of course, we do not need to adopt a *monistic* axiology, i.e the claim that there is only one final value (Robbins, 2013). Final value can be monistic or pluralist, and it can be relate to instrumental values along many different dimensions and descriptive hierarchies. All that is needed for indirect normativity, and consequently AGI alignment, to be viable is that we can *in principle* close the gap between the evaluative structure of humanity (or of the coherent extrapolated volition of humanity), and the evaluative structure of AGI. If we can correctly implement a process for closing such a gap, then the pro-moral behaviour of an AGI will increase as its cognitive capacities improve.

Limitations

AGI alignment, and AGI safety more broadly, is a vast area of research, which is of deep importance. I am indebted to the continued efforts of the those who are working towards mitigating these risks, especially the two giants of conceptual AGI safety work, Eliezer Yudkowsky and Nick Bostrom, whose work I have tried to do justice. There is much more to this subject than I can hope to cover here. I have not attempted to provide an account of final values in terms of the actual ends to which we are aiming with AGI alignment—this has been intentional, as my arguments apply no matter what kind of final value we are considering, whether it is a collective conception (as is the case for Yudkowsky's CEV), or an individual (as is the case for Christiano's formalisation of indirect normativity). There are also ethical concerns with the focus on human final values as the axiological ends of alignment.

In that my focus has been on the unviability of direct specification I have not engaged in an in-depth defence of the viability of indirect normativity, which generates its own web of concerns. Of special note is the concept of *corrigibility*—the design of systems that do not converge on supergoal preservation, i.e systems that are tolerant to errors in specification (Soares et al, 2015), which would include errors in any specified processes for value updating. Whether a goal-directed system can exist without a fixed reference point which specifies the process for updating is central to the viability of true instantiations of indirect normativity. Future conceptual work is needed to ground, or perhaps undermine, this possibility.

Conclusion

In this thesis I have argued that AGI development is a source for concern which bears the potential to be an existential risk, and that we should not attempt to mitigate this risk by

designing and implementing an explicit AGI value structure. The project of AGI alignment must be set against the backdrop of the possibility of superintelligence; if an intelligence explosion occurs, there is a significant chance that holes in our specification will be torn open. Several researchers have already identified this problem, noting that value specification, and the construction of an architecture of morality, have been historically ineffective. All the more so when the agent whose values we specify is entirely artificial, without pre-embedded evolutionary or socially moulded structures that can shunt specification into the right interpretative channels. Bostrom (2014) has defined a family of methods for value alignment he calls "indirect normativity" which, if successful, can offload a significant portion of this work to the artificial agent itself. If this is properly implemented, value specifications are taken by the agent as mere approximations; an initial dynamic which can be updated. In such a case an intelligence explosion will shrink, rather than tear open, the holes in specification.

I have here built on those who reject specification by constructing a conceptual account of the holes in specification attempts, and the consequent unviability of direct specification. This unviability, I have shown, is not just a consequence of our observations, past and present, of the failings of attempts at formulating value specifications. It runs deeper than that, as a consequence of the holes in specification and their immutability. These holes, I have argued, can be understood as the inevitable corruption of meaning when concepts are translated into new frames. This is a result from the cognitive and axiological organisation of moral concepts: descriptive abstraction and axiological reference to final values both involve hierarchical cognitive structures. The translation between different hierarchies and between levels of these hierarchies results in new semiotic structures, which is the conceptual core of the holes in specification.

Issues of framing, translation, and mapping between semiotic spaces are ubiquitous. In certain domains they cause more problems than elsewhere. Nowhere is this truer than in AGI value specification. Framing errors between humans can cause miscommunications and misunderstandings, but as humans we share enough of our semantic and axiological backdrop that these misunderstandings are often made clear to all involved as soon as they carry implications that seem obviously bizarre. We have enough of a shared reference frame, enough interpretive and normative "common sense", that some degree of alignment is almost inevitable. If this were not the case, then miscommunication would be a much more profoundly catastrophic issue. AGIs, on the other hand, have no axiological reference frame except what we give them; no attractor points in the human ethical space.

Direct specification is unviable in that it does not succeed in the task of *true* AGI alignment—the alignment of our final values with the AGI's final values. This stringent standard of

alignment is crucial for AGIs. Even subtle points of departure between human and AGI values have the potential to magnify themselves, particularly given the alien conceptual spaces that AGI, particularly superintelligence, will be moving through. How much faith should we have in our ethical concepts and frameworks as they are now, and in our ability to translate them into forms suitable for AGI programming? Superintelligence has been described as "runaway AI" (Tucker, 2006). This term is appropriate: the convergent goals of supergoal preservation and self-preservation ensure that once an AGI has the capacity to evade our value-altering grasp, it will. Any holes that are already present in value specification will be set in silicone, unless we allow for those holes to be closed.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology, 7*(3): 149-155.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016). Concrete problems in AI safety. *arXiv:1606.06565*.

Andersen, H., Barker, P., Chen, X. (2006). *The cognitive structure of scientific revolutions*. Cambridge University Press.

Anderson, M., & Anderson, S. L. (2007). The status of machine ethics: a report from the AAAI Symposium. *Minds and Machines, 17*(1): 1-10.

Anderson, M., & Anderson, S. L. (eds.). (2011). *Machine ethics*. NY: Cambridge University Press.

Aristotle. (1908). *The Nicomachean ethics*. Translated by W. D. Ross. Accessed at http://classics.mit.edu/Aristotle/nicomachaen.html Originally written approximately 340 BCE.

Arkoudas, K., Bringsjord, S., Bello, P. (2005). Toward ethical robots via mechanized deontic logic. In *AAAI fall symposium on machine ethics*: 7-23.

Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines, 22*(4): 299-324.

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics and Philosophy 16*: 247-266.

Asimov, I. (1942). Runaround. *Astounding Science Fiction, 29*(1), 94-103.

Barrett, A. M., & Baum, S. D. (2017). A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence, 29*(2): 397-414.

Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In Lehrer, A., Kittay, E.F., Lehrer, R. (eds). *Frames, fields, and contrasts: new essays in semantic and lexical organization*. NY: Routledge: 21-74.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron 76*:695-711.

Bogosian, K. (2017). Implementation of Moral Uncertainty in Intelligent Machines. *Minds and Machines, 27*(4): 591-608.

Bonissone, P. P. (2010). Soft computing: A continuously evolving concept. *International Journal of computational Intelligence Systems, 3*(2): 237-248.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*: 277-284.

Bostrom, N., Yudkowsky, E. (2011). The ethics of artificial intelligence. In Ramsey, W., Frankish, K (eds). *Cambridge Handbook of Artificial Intelligence*. UK: Cambridge University Press: 316-335.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines, 22*(2): 71-85.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. UK: Oxford University Press.

Bostrom, N., Dafoe, A., Flynn, C. (2016). Policy desiderata in the development of machine superintelligence. *Future of Humanity Institute*, University of Oxford.

Bourget, D., Chalmers, D. J. (2014). What do philosophers believe?. *Philosophical studies, 170*(3): 465-500.

Bringsjord, S., Arkoudas, K., Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems, 21*(4): 38-44.

Bringsjord, S., Bringsjord, A., & Bello, P. (2012). Belief in the singularity is fideistic. In *Singularity Hypotheses*. Berlin: Springer: 395-412.

Brosch, T., Sander, D. (eds). (2016). *Handbook of value: perspectives from economics, neuroscience, philosophy, psychology, and sociology*. NY: Oxford University Press.

Buckner, C. (2018). Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese 195*(12):5339–5372.

Burkart, J., Schubiger, M., & Van Schaik, C. (2017). The evolution of general intelligence. *Behavioral and Brain Sciences 40*. E195.

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2018). Motivations and risks of machine ethics. *Proceedings of the IEEE, 107*(3): 562-574.

Chajewska, U., Koller, D., Parr, R. (2000). Making rational decisions using adaptive utility elicitation. *Proceedings of the Seventeenth National Conference on Artificial Intelligence*: 363-369.

Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies 17*: 7-65.

Christiano, P. (2012). *A formalization of indirect normativity*. From https://ai-alignment.com/a-formalization-of-indirect-normativity-7e44db640160.

Clark, A. (2002). Local associations and global reason: Fodor's frame problem and second-order search. *Cognitive Science Quarterly 2*(2):115-140.

Coulson, S. (2001). *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. NY: Cambridge University Press.

Crnkovic, G. D., & Çürüklü, B. (2012). Robots: ethical by design. *Ethics and Information Technology, 14*(1): 61-71.

Dancy, J. (1983). Ethical particularism and morally relevant properties. *Mind, 92*(368): 530-547.

Dancy, J. (1993). *Moral reasons*. Oxford: Blackwell.

Dancy, J. (2004). *Ethics without principles*. Oxford: Clarendon.

Datteri, E. (2013). Predicting the long-term effects of human-robot interaction: A reflection on responsibility in medical robotics. *Science and engineering ethics 19*(1): 139-160.

Dennett, D. (2005). Cognitive wheels: the frame problem of AI. *Language and Thought 3*: 217-233.

Dewey, D. (2011). Learning what to value. In Schmidhuber, J., Thórisson, K. R., Looks, M. (eds). *Artificial General Intelligence: 4th International Conference Proceedings*: 309–314.

Dreyfus, H. L. (1979). *What computers can't do: The limits of artificial intelligence*. New York: Harper & Row.

Dumoulin, S. O. (2017). Layers of Neuroscience. *Neuron 96*:1205-1206.

Eckersley, P. (2019). Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function). *arXiv*: https://arxiv.org/pdf/1901.00064.pdf.

Eden, A. H., Moor, J. H., Søraker, J. H., Steinhart, E. (eds.) (2012). *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Berlin: Springer.

Elstein, D. Y., & Hurka, T. (2009). From thick to thin: Two moral reduction plans. Canadian Journal of Philosophy, 39(4), 515-535.

Everitt, T., Lea, G., Hutter, M. (2018). AGI Safety Literature Review. *International Joint Conference on Artificial Intelligence (IJCAI)*.

Feldman Barrett, L. (2018). *How emotions are made: the secret life of the brain*. London: Pan.

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the morning calm* (ed. The Linguistic Society of Korea). Seoul: Hanshin: 111-137.

Frey, R. G. (2013). Act-Utilitarianism. In LaFollette, H., & Persson, I. (eds.). *The Blackwell guide to ethical theory*. John Wiley & Sons.

Gamerschlag, T., Gerland, D., Osswald, R., Petersen, W. (eds). (2015). *Meaning, Frames, and Conceptual Representation*. Düsseldorf: Düsseldorf University Press.

Gilbert, M. (2001). Collective preferences, obligations, and rational choice. *Economics & Philosophy, 17*(1): 109-119.

Goertzel, B., Pennachin, C. (eds). (2007). *Artificial General Intelligence*. Berlin: Springer.

Goldin, D., Wegner, P. (2008). The interactive nature of computing: Refuting the strong Church–Turing thesis. *Minds and Machines, 18*(1): 17-38.

Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. Advances in computers 6: 31-88.

Goodall, N. J. (2014). Vehicle automation and the duty to act. In *Proceedings of the 21st world congress on intelligent transport systems*: 7-11.

Grim, P., Denis, P. S., Kokalis, T. (2004). Information and meaning: Use-based models in arrays of neural nets. *Minds and Machines, 14*(1): 43-66.

Guarini, M. (2010). Particularism, analogy, and moral cognition. Minds and Machines, 20(3), 385-422.

Gunkel, D. J. (2017). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology*: 1-14.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in neural information processing systems 29*: 3909-3917.

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., & Dragan, A. (2017). Inverse reward design. *Advances in neural information processing systems 30*:6765-6774.

Hakli, R., Mäkelä, P. (2019). Moral Responsibility of Robots and Hybrid Agents. *The Monist, 102*(2): 259-275.

Heckerman, D., Breese, J. S., Horvitz, E. J. (2013). The compilation of decision models. *arXiv*:1304.1510. Originally in *Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence 1989*.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: an ethical analysis. *Science and engineering ethics 21*(3): 619-630.

Hill, T. E., & Hill Jr, T. E. (2000). *Respect, pluralism, and justice: Kantian perspectives*. Oxford: Clarendon.

Holekamp, K. E., Miikkulainen, R. (2017). The evolution of general intelligence in all animals and machines. *Behavioral and Brain Sciences 40*.

Hooker, B., Little, M. O. (eds). (2000). *Moral particularism*. Oxford: Clarendon.

Hooker, J. N., & Kim, T. W. N. (2018). Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*: 130-136.

Horgan, T., Timmons, M. (2009). What does the frame problem tell us about moral normativity?. *Ethical Theory and Moral Practice, 12*(1): 25-51.

Hutter, M. (2007). Universal algorithmic intelligence: A mathematical top→ down approach. In Goertzel, B., Pennachin, C. *Artificial general intelligence*. Berlin: Springer: 227-290.

Inthorn, J., Tabacchi, M. E., & Seising, R. (2015). Having the final say: Machine support of ethical decisions of doctors. In *Machine Medical Ethics*. Springer, Cham: 181-206.

Jiga-Boy, G. M., Maio, G. R., Haddock, G., Tapper, K. (2016). Values and behavior. In Brosch, T., Sander, D. (eds). *Handbook of value: perspectives from economics, neuroscience, philosophy, psychology, and sociology*. NY: Oxford University Press: 243-262.

Kant, I. (2012). *Groundwork of the metaphysics of morals*. Mary Gregor, M., Jens Timmermann, J. (Trans.). NY: Cambridge University Press. Originally published 1785

Kawall, J. (2006). On the moral epistemology of ideal observer theories. *Ethical Theory and Moral Practice, 9*(3): 359-374.

Kiebel, S. J., Daunizeau, J., Friston, K. J. (2009). Perception and hierarchical dynamics. *Frontiers in neuroinformatics 3*: 20.

Kim, T. W., Donaldson, T., Hooker, J. (2018). Mimetic vs Anchored Value Alignment in Artificial Intelligence. *arXiv*:1810.11116.

King, D. (1996). Is the human mind a Turing machine?. *Synthese, 108*(3): 379-389.

Korsgaard, C. M. (1983). Two Distinctions in Goodness. *The Philosophical Review 92*(2): 169-19.

Kuleshov, V., Schrijvers, O. (2015). Inverse game theory. *WINE 2015 Proceedings of the 11th International Conference on Web and Internet Economics*. Berlin: Springer: 413-427.

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and machines, 17*(4): 391-444.

Leveringhaus, A. (2018). What's So Bad About Killer Robots?. *Journal of Applied Philosophy, 35*(2): 341-358.

List, C. (2014). Three kinds of collective attitudes. *Erkenntnis, 79*(9): 1601-1622.

List, C. (2016). Levels: descriptive, explanatory, and ontological. *Noûs*. Accessed at Wiley Online Library: https://onlinelibrary.wiley.com/doi/abs/10.1111/nous.12241

Little, M. O. (2000). Moral generalities revisited. In Hooker, B., Little, M. O. (eds). *Moral particularism*. Oxford: Clarendon. 276-311.

Lokhorst, G. J., Van Den Hoven, J. (2011). Responsibility for military robots. In Lin, P., Abney, K., Bekey, G. A. (eds.) *Robot ethics: The ethical and social implications of robotics*. MIT Press: 145-156.

Mangalaraj, G., Nerur, S., Mahapatra, R., & Price, K. H. (2014). Distributed Cognition in Software Design: An Experimental Investigation of the Role of Design Patterns and Collaboration. MIS Quarterly, 38(1): 249-274.

Martins, M. D., Gingras, B., Puig-Waldmueller, E., Fitch, W. T. (2017). Cognitive representation of “musical fractals”: Processing hierarchy and recursion in the auditory domain. *Cognition 161*:31-45.

McAskill, W. (2014). *Normative Uncertainty*. Dissertation, University of Oxford.

McCarthy, J. (1969). Hayes. PJ: Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence 4*: 463-502.

McCauly, L. (2007). AI Armageddon and the Three Laws of Robotics. *Ethics and Information Technology 9*(2): 153-164.

McNaughton, D., Rawling, P. (2000). Unprincipled ethics. In Hooker, B., Little, M. O. (eds). *Moral particularism*. Oxford: Clarendon. 256-275.

Mindermann, S., Shah, R., Gleave, A., & Hadfield-Menell, D. (2018). Active Inverse Reward Design. *arXiv*:1809.03060.

Minsky, M. (1975). A Framework for Representing Knowledge. In Winston, P. (ed). *The Psychology of Computer Vision*. McGraw-Hill.

Moore, G. E. (1959). *Principia Ethica*. NY: Cambridge University Press. Originally published 1903.

Moor, J. H. (2009). Four kinds of ethical robots. *Philosophy today 72*: 12-14.

Moor, J. H. (2011). The nature, importance, and difficulty of machine ethics. In Anderson, M., & Anderson, S. L. (eds.). *Machine ethics*: 13-20.

Muehlhauser, L., & Helm, L. (2012). The singularity and machine ethics. In *Singularity Hypotheses*. Berlin: Springer: 101-126

Muehlhauser, L., Bostrom, N. (2014). Why we need friendly AI. *Think, 13*(36): 41-47.

Natarajan, S., Kunapuli, G., Judah, K., Tadepalli, P., Kersting, K., Shavlik, J. (2010). Multi-agent inverse reinforcement learning. *International Conference on Machine Learning and Applications*.

Nozick, R. (1974). Anarchy, state, and utopia (Vol. 5038). New York: Basic Books.

Nyholm, S. (2018). Attributing agency to automated systems: reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics 24*(4): 1201-1219.

Oesterheld, C. (2016). Backup utility functions as a fail-safe AI technique. *Foundational Research Institute*. From https://foundational-research.org/files/backup-utility-functions.pdf.

Olson, J. (2004). Intrinsicalism and conditionalism about final value. Ethical Theory and Moral Practice, 7(1), 31-52.

Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness, 5*(02): 105-129.

Penrose, R. (1994). *Shadows of the Mind*. Oxford: Oxford University Press.

Plebe, A., & Perconti, P. (2012). The slowdown hypothesis. In *Singularity hypotheses*. Berlin: Springer: 349-365.

Prinzing, M. (2017). Friendly superintelligent AI: All you need is love. In *3rd Conference on Philosophy and Theory of Artificial Intelligence*. Springer, Cham: 288-301.\

Robbins, J. (2013). Monism, pluralism and the structure of value relations. *HAU Journal of Ethnographic Theory 3*(1):99.

Roberts, D. (2011). Shapelessness and the Thick. *Ethics, 121*(3), 489-520.

Rumsfeld, D. (2002). Press Conference, accessed at https://www.nato.int/docu/speech/2002/s020606g.htm.

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: a modern approach* (3rd edition). Pearson.

Saitta L., Zucker J. D. (2013). *Abstraction in Artificial Intelligence and Complex Systems*. NY: Springer.

Scarcello, O. (2018). On the role of normative hierarchies in constitutional reasoning: a survey of some paradigmatic cases. *Ratio Juris 31*(3):346-363.

Schurz, G. & Votsis, I. (2007). A Preliminary Application of Frame-Theory to the Philosophy of Science: The Phlogiston-Oxygen Case. *Dusseldorf CPT conference*.

Schurz, G. & Votsis, I. (2012). A frame-theoretic analysis of two rival conceptions of heat. *Studies in History and Philosophy of Science 43*: 105-114.

Sepulcre, J., Sabuncu, M. R., Yeo, T. B., Liu, H., Johnson, K. A. (2012). Stepwise connectivity of the modal cortex reveals the multimodal organization of the human brain. *The Journal of Neuroscience 32*(31):10649 –1066.

Shulman, C. (2010). *Omohundro's "basic AI drives" and catastrophic risks*. Manuscript.

Soares, N., Fallenstein, B. (2015). Two attempts to formalize counterpossible reasoning in deterministic settings. In *International Conference on Artificial General Intelligence* (pp. 156-165). Springer, Cham.

Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Accessed at https://intelligence.org/files/Corrigibility.pdf.

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy, 24*(1): 62-77.

Sullins, J. P. (2005). Ethics and artificial life: From modeling to moral agents. *Ethics and Information technology, 7*(3): 139.

Sullins, J. P. (2011). When is a robot a moral agent. In Anderson, M., & Anderson, S. L. (eds.). *Machine ethics*: 151-160.

Sun, R. (1996). Hybrid Connectionist-Symbolic Modules: A Report from the IJCAI-95 Workshop on Connectionist-Symbolic Integration. *AI Magazine, 17*(2):99.

Tappolet, C., Rossi, M. (2016). What is value? Where does it come from? A philosophical perspective. In Brosch, T., Sander, D. (eds). *Handbook of value: perspectives from economics, neuroscience, philosophy, psychology, and sociology*. NY: Oxford University Press. 3-22.

Taylor, J., Yudkowsky, E., LaVictoire, P., Critch, A. (2016). Alignment for advanced machine learning systems. Machine Intelligence Research Institute.

Timmons, M. (2017). The categorical imperative and universalizability. In *Significance and system: essays on Kant's ethics*. Oxford Scholarship Online.

Thiele, L. P. (2006). *The Heart of Judgement*. NY: Cambridge University Press.

Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI & Society, 22*(4): 495-521.

Torres, P. (2016). Agential Risks: A Comprehensive Introduction. *Journal of Evolution and Technology 26*(2):31-47.

Torres, P. (2018). Superintelligence and the future of governance: on prioritizing the control problem at the end of history. In *Artificial Intelligence Safety and Security*. Chapman and Hall: 357-374.

Tsu, P. S. H. (2013). Shapelessness and predication supervenience: a limited defense of shapeless moral particularism. *Philosophical Studies, 166*(1), 51-67.

Tucker, M. (2019). From an axiological standpoint. Ratio, 32(2), 131-138.

Tucker, P. (2006). The singularity and human destiny. *The Futurist, 40*(2), 3.

Van Den Hoven, J., Lokhorst, G. J. (2002). Deontic Logic and Computer-Supported Computer Ethics. *Metaphilosophy, 33*(3): 376-386.

Van Rysewyk, S. P., Pontier, M. (2014). *Machine medical ethics*. Springer.

Väyrynen, P. (2014). Shapelessness in Context. *Noûs 48*(3):573-593.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wallach, W., Allen, C., Franklin, S. (2011). Consciousness and ethics: Artificially conscious moral agents. *International Journal of Machine Consciousness, 3*(01): 177-192.

Waser, M. R. (2008). Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence. *AAAI Fall Symposium: Biologically Inspired Cognitive Architectures*: 195-200.

Waser, M. R. (2009). What is artificial general intelligence? Clarifying the goal for engineering and evaluation. In *Proceedings of the 2nd Conference on Artificial General Intelligence*. Atlantis Press.

Wermter S., Sun R. (1998). An Overview of Hybrid Neural Systems. In Wermter S., Sun R. (eds). *Hybrid Neural Systems.* Lecture Notes in Computer Science (1778). Berlin: Springer.

Whitby, B. (2015). Automating medicine the ethical way. In *Machine Medical Ethics*. Springer, Cham: 223-232.

Wittgenstein, L. (1969). *Philosophical investigations* (3rd ed.) Oxford: Blackwell. Originally published 1953.

Wu, B., Feng, Y., Zheng, H. (2014). Model-based Bayesian reinforcement learning in factored Markov decision process. *Journal of Computers, 9*(4): 845-850.

Yampolskiy, R. (2016). Taxonomy of pathways to dangerous Artificial Intelligence. *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence AI, Ethics, and Society: Technical Report*: 143-148.

Yudkowsky, E. (2004). Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence.*

Yudkowsky, E. (2008). AI as positive and negative factor in global risk. In Bostrom, N., Cirkovic, M. M. (eds). *Global catastrophic risks*. NY: Oxford University Press: 308-345.

Yudkowsky, E. (2011). Complex value systems in friendly AI. *International Conference on Artificial General Intelligence*. Berlin: Springer: 388-393.

Yudkowsky, E. (2016). The AI Alignment Problem: Why it is Hard, and Where to Start. *Symbolic Systems Distinguished Speaker*.

Zahn, R., Moll, J., Paiva, M., Garrido, G., Krueger, F., Huey, E. D., Grafman, J. (2009). The neural basis of human social values: evidence from functional MRI. *Cerebral Cortex 19*(2): 276–283.