The Empowerment of Others as a solution to the AI Control Problem

Sam Brown

January 7, 2022

AI alignment, or the AI control problem, concerns the question of how to build systems which successfully aid their creator without inadvertently causing harm. This essay discusses some of the difficulties of this problem in the context of possible "superintelligent" systems: ones with intelligence surpassing that of human minds, who may develop arbitrarily-high levels of intelligence and power[1]. I then introduce an outwards-looking version of the existing concept of Empowerment, and examine its suitability as a target of AI alignment. I argue that this version of Empowerment provides a route to human flourishing while avoiding the risks of piecemeal target construction and minimising paternalism towards future generations.

Today, computer systems can perform at levels approaching and surpassing human experts, at a variety of specific reasoning tasks from games such as chess and Go (see e.g. AlphaGo Zero [Silver et al., 2017]) to the production of mathematical proofs (see e.g. the Kepler conjecture [Hales, 2005] and the four colour theorem [Appel and Haken, 1989]). Current research is successfully creating increasingly general-purpose systems [Schmid et al., 2021] which are able to complete "zero-shot" tasks, those at which they have not been explicitly trained by their creator[2].

Current-day AI systems are not superintelligent, though they may outperform humans at specific tasks[3]. However, the possibility remains that

[1]Here we follow Legg and Hutter [2007] and define 'intelligence' as "an agent's ability to achieve goals in a wide range of environments".

[2]See for example GPT-3 [Brown et al., 2020], which — having been trained to suggest predictive completions from text prompts — can complete tasks such as language translation [§3.3], arithmetic [§3.9.1] and answering SAT tests [§3.9.3].

[3]Nor does superintelligent need to arise from machine-learning systems: Bostrom [2014]

superintelligent agents may come to exist in the future, and that humans may have a hand in their creation. A survey of AI experts in 2012/2013 [Müller and Bostrom, 2016] found a median expectation of superintelligence within this century.

The term 'target' of AI alignment refers to the desired outcome of an agent's behaviour. Often this is known in an informal sense ("clean a room", "make paperclips", or "increase human satisfaction") but is difficult to define formally and rigorously to a machine that has no preconception of "common sense".

Machine learning (ML) is often a much more effective problem-solving approach than traditional methods of programming, but hands much of the control of *how* the problem is solved to the program. It differs from traditional methods in that the problem-solving rules which the program/agent follows are not explicitly defined by the programmer. Instead, the programmer provides explicit rules on *how* to learn, defines a desired output, and provides relevant data; the program/agent then learns its own rules from the data provided. An example of a traditional technique to distinguish photos of dogs from photos of cats might attempt to explicitly measure features such as "pointiness of ears" and make a judgement predefined by the programmer. In comparison, a machine-learning approach might provide labelled examples of both cat photos and dog photos, and provide the program with the ability to extract basic visual features from images (e.g. line-detection), so that the ML program can use the data provided to learn explicit "cat-or-dog" rules itself.

In many ML techniques there is some "objective function" whose output is to be maximised by the agent. A core challenge then is to represent the problem in the form of a mathematical function whose output is a value which can be maximised.

This essay explores the limitations of various approaches to constructing an objective function aligned with human values, and proposes a candidate which aims to avoid these. First, in section 1, we cover some basic assumptions, concerning properties that superintelligent agents might have. Then, sections 2 and 3 discuss two approaches to creating a target for AI alignment: negative definitions (what we want an agent *not* to do) and positive definitions (what we want an agent to actively work towards). I argue that

suggests various paths to superintelligence, including genetic engineering, brain emulation, and nootropics.

negative definitions are insufficient for reliable AI safety, and that a unified positive target of alignment is necessary. Section 3.1 discusses various ways of aggregating values and agreeing on actions. Section 3.2 discusses moral issues of paternalistically taking actions on behalf of others, while section 3.3 argues that *some* decision *must* be made. Section 4 proposes a positive target for AI alignment, based on the idea of personal liberty. Section 4.2 discusses the concept of Empowerment from information theory, which is used here in the context of personal liberty. Section 4.3 introduces the idea of extending the subjective concept of Empowerment outwards, to others, and 4.4 suggestions the emulation of the minds of others as a mechanism for this extension. Finally, section 4.5 mentions limitations and challenges of this proposed target for AI alignment, such as the remaining choice of beneficiary of this liberty, and the challenge of weighting options so that trivial liberties do not overwhelm ones which people find meaningful.

In sum, I aim to develop a way that would allow a superintelligent entity to model and seek the empowerment of people, and therefore an equitable shared agency and liberty.

1 Some properties of superintelligent agents

Superintelligent agents are expected to pursue general-purpose instrumental goals [Bostrom, 2014] [Russell, 2019]: ones which make most end-goals more achievable. These assumptions may seem extreme and pessimistic, and are not universally accepted, but the purpose of this essay is not to defend them. Rather, these proposed properties of superintelligent agents outline the scope of the problem for which I argue a positive target of alignment is necessary.

Intelligence explosion For an entity to be superintelligent requires only that it be more generally intelligent than humans. But there is a general view that once an entity surpasses that threshold, assuming that it will be able to modify itself, it will proceed to rapidly increase in intelligence without an obvious limit. This follows from the following feedback loop: if an entity with the intelligence of humans (H) is able to create an entity with superhuman intelligence ($H + a$), then an entity with superhuman intelligence ($H + a$) could feasibly produce an entity more intelligent still ($H + a + b$), and so on. This essay assumes that a superintelligent entity could become arbitrarily intelligent.

Control of the physical world An agent may begin with some measure of control over the physical world, e.g. control over a factory. If not, in order for an initially-isolated agent to bootstrap physical control it has been suggested that a superior intelligence would be able to convince its operators (who *do* have access to the outside world) to move in the world on its behalf. This essay assumes that the agent can gain arbitrary levels of control over the physical world.

Resistance to value update An agent's goal is threatened by the prospect of its motivating values being changed, and so it would oppose such an alteration. For example, if the objective function of Bostrom's infamous paperclip maximiser [2003] were changed to no longer value paperclips above all else, its future would contain fewer paperclips, and so it would resist such a value update. This essay therefore assumes that any "desire for update" must be preemptively designed and encoded[4].

In summary, the fear is that humans may develop a superintelligent system which develops arbitrarily-high levels of intelligence and control over the physical world, and that resists new input or control from humans. Such a system must be created predisposed to agree to our human values. In the following sections I consider two approaches to defining our values, so that they can be encoded within the system: negative and positive.

2 Negative definitions

In this essay, I am primarily concerned with *positive* definitions. That is to say: this essay explores the question of what behaviour we would want a superintelligent agent to actively engage in.

It is possible to come up with *negative* definitions of the target of AI alignment: we can agree on ways we want an agent *not* to behave. The paperclip maximiser in Bostrom's [2003] thought experiment - in the pursuit of creating as many paperclips as possible - uses all the world's available resources to that end, including the atoms which until then belonged in human

[4]A special case is the "off-switch" problem: an AI may rationally try to seek self-preservation by circumventing its own "off-switch" (e.g. by making an external copy of itself to pursue its goals after the death of the original) unless it is otherwise predisposed [Hadfield-Menell et al., 2016].

bodies. One can illustrate the point without extreme thought experiments: Amodei et al. [2016] use the more moderate language of "accidents" to describe unintended and harmful side-effects, giving practical examples such as a cleaning robot knocking over a vase in order to clean faster.

The distinction between negative and positive approaches has been discussed by Gabriel [2020], who terms them 'minimalist' and 'maximalist' conceptions of value alignment respectively. Gabriel suggests that the cost of a minimalist approach is one of opportunity cost: that such agents may be "safe and reliable but still a long way from what is best" and that "we may ultimately need to move beyond minimalist conceptions if we are going to produce fully aligned AI". This hinges on the assumption that a minimalist conception will be sufficient as a safe first step, which can later be improved.

I disagree: definitions of undesirable outcomes are not sufficient, and it is dangerous to rely on them exclusively. It is impossible to comprehensively predict and list undesired side-effects. Approaches by Amodei et al. [2016] give various ways to transferably generalise undesired behaviours ("don't knock objects over" is more general and transferable than listing specific objects), and methods of "regularising" behaviour (creating secondary tie-breaking rules e.g. by creating "low impact agents" which avoid unwarranted power-seeking behaviour). However, these approaches merely *limit* predicted side-effects in an attempt to bound incidental harm, rather than avoiding them comprehensively and reliably.

2.1 Unpredictability of trade-offs

Methods of optimisation by definition seek to find extreme behaviours. The calculus of variations, for example, is concerned with functions' extrema. A linear programming optimization which is not well-bounded will drive towards an extreme solution state so that the output of the objective function increases unbounded. Any human value to which the objective function is ambivalent will be sacrificed in the pursuit of even very minor gains in whatever the agent *does* value.

'Reward hacking' is a common feature of machine learning approaches such as reinforcement learning. While evolving virtual creatures, Sims [1994] sees many behaviours which violate the spirit, but not the letter, of the coded rules. For example, virtual creatures judged on how fast they can move — with the hope that they would learn to walk or swim — instead evolved behaviours such as oscillating rapidly back and forth, or growing very fast and

then falling over. In an emulated 'hide-and-seek' environment, Baker et al. [2019] discover that their agents find ways to exploit the built environment and physics engine in unintended ways which "completely surprised"[p.22] the authors. While these approaches are often comical, their unexpectedness and ingenuity are exactly the point of these methods of machine learning: if we knew the rules in advance we could code them in explicitly.

The unpredictability of the reward-maximising behaviour of machine-learning agents makes a focus on avoiding negative side-effects insufficient and unsafe. As an analogy, negating negative definitions corresponds to constructing flood defences piecemeal: the joins between the components must be watertight. The water will find any gaps and — once through them — will go where it wills.

In contrast to e.g. Gabriel [2020] who implicitly describes the costs of a 'minimalist' conception of alignment as one only of opportunity cost, I argue that "good enough" is not good enough. A system of alignment constructed only from negative definitions (even from a set of generalised definitions) may well be leaky, and an agent with optimization at its core will find the gaps. The properties of superintelligent agents described in §1 mean that we may not have the opportunity to rely on iteratively correcting our instructions as we realise their shortcomings. Instead, we must encode what it is we want a superintelligent agent to work *towards*.

3 Positive definitions

Rather than defining what we *don't* want, I argue it is better to define what we *do* want, and have the AI's behaviour gravitate *towards*, rather than away from, that definition.

One might well question this distinction. After all, if we're trying to maximise the output of some objective function V, is that not the same as minimising some opposite function $-V$? And of course this is true. My core distinction is really about the *unity* of the objective function, and in part the fundamentality of it. Most choices of negative definitions refer to partial and instrumental goals, and so do not mesh well with each other. A comprehensive definition of terminal disvalue may well be suitable as a term for an objective function to minimise, but I am not aware of any such definition.

The question of whether AI alignment's target can be positively defined

has two parts: the definition of the target to a level that is unambiguous and comprehensible to humans, and the communication of that definition to a computer system through some formal encoding.

Let us first focus on the identification of a target. When we consider that the purpose of a superintelligent agent is to maximise the value of this target, and we assume that it can become powerful enough to achieve any physically-possible goal, we see that the question of AI alignment is one of utopia: this desired target is nothing other than a metric of human flourishing[5].

But the answer to the question "what does human flourishing look like?" is controversial. Many people have conflicting opinions; politics is competitive. Rawls [1993] writes of "reasonable pluralism": a plurality of reasonable yet incompatible doctrines.

Perhaps we can take hope in the notion that these political disagreements are about factual beliefs, rather than about values or desires. When one person opposes another's immigration policy, say, we may hope that the disagreement lies in their beliefs about the real-world effects of the policy, and the extent to which it will be effective in increasing human flourishing, rather than disagreement on fundamental values (though see Haidt's research on differing "moral tastes" [2012]).

Or perhaps such disagreements can always be understood to be subjective, in the manner of how to honour one's dead (e.g. cremation vs necrophagy), which meals are forbidden, or which side of the road one should drive on. The importance is less on the particular choice made, but rather that *some* choice is made and agreed on within a society.

We need, then, to work towards the crux of disagreement, if we hope to find universally agreed values.

However, even cooperative investigatory fields, such as academic research into axiology, ethics and meta-ethics, yields significant disagreement. In a survey of modern philosophers, Bourget and Chalmers [2013] find that philosophers are undecided even between broad strokes, split between deontology (26%), consequentialism (24%), and virtue ethics (18%), with the remaining 32% a miscellaneous "other". Half of respondents (56%) were

[5]One might ask about, for example, animal flourishing. But since humans are the ones hypothetically creating this superintelligent AI, it would seem that animal flourishing should be prioritised exactly as much as humans desire. This raises a distinction between an *aligned* AI, and a *moral* AI: it is conceivable that an AI could be successfully aligned to the values of an evil creator. This essay aims to find a target which is resistant to capture by an individual.

moral realists, though only half of the rest (28%) were moral anti-realists. Even within a single branch of philosophy, say utilitarianism, there are generally a multitude of potential denominations with axioms to fit a variety of intuitions.

What, then, can we agree on?

3.1 How to handle disagreement

Many political systems, from autocracy to aristocracy to democracy, define ways of aggregating individuals' choices of actions, or choices of high-level desires (e.g. taxation policies are a single distillation of the preferences of many). Are there analogous ways of aggregating the fundamental values found at the end of a collaborative exploration of our disagreements?

A moral realist, intent on aligning a superintelligent agent to absolute morality, might find the question misguided. After all, if there is a moral truth, it exists regardless of whether anyone believes it, let alone agrees on it. A moral realist, then, who is convinced of some moral truth, might wish to encode it in a superintelligent entity regardless of the opinions of others. Claims of the unrecognisability of the good in their chosen end (e.g. a universe filled with hedonium [Bostrom, 2014]) would be irrelevant.

If, instead, we think that the views of others are important, we must find a way to resolve conflict, which as we have seen seems a fundamental part of life [Krahulik and Holkins, 2004].

Ways to navigate disagreement while avoiding coercion in a liberal state generally rely on underlying agreement between the disagreeing parties: there is some deeper truth or framework on which they *do* agree. For example, the disputants might agree to abide by the decision of a mutually-agreed-upon third party.

But this underlying agreement is not always possible, and there are sometimes conflicts which demand resolution. It is key to note, here, that where violent conflict arises, the disputants are already prepared to violate each other's will (or else there would be an underlying agreement: that another's will is inviolable). In a situation where such fundamental conflict already exists, hard paternalism is inevitable.

3.2 Paternalism

If we presume to make decisions on behalf of other people, even knowing that they may not immediately agree with us, we may expect accusations of paternalism.

Paternalism can be succinctly defined as "benevolent coercion"[6]. Distinctions are made between 'hard' paternalism and 'soft' paternalism, on the basis of whether the behaviour being interfered with is known to be voluntary, or whether it may be involuntary. In Mill's example [1974] of a person about to cross a rotten bridge, it is soft paternalism to stop them to check that they know the bridge is unsafe, but hard paternalism to learn that the person intends to take their own life jumping from the bridge and forcibly prevent them.

Generally, some level of paternalism is socially accepted, and even desired (for example, compulsory school curricula, drug information campaigns and seatbelt mandates) [Begon, 2016]. It may be argued that all human laws are paternalistic[7].

Paternalism is defended by the opinion that one is better informed than the person whose agency is being impinged upon. Where this is truly the case (as in the soft paternalism of a child prevented from running into traffic), it seems best to embrace this truth and shoulder the attendant responsibility.

But the case in AI alignment is not so straightforward. Instead of directly impinging on the agency of future generations (who we do not assume are less informed or rational), we are forced to make a decision about how to direct an agent which may well become better informed than any of our descendants. We presume to direct Plato's philosopher-king, "who, as philosopher, would know what to do and, as king, would be able to do it" [Skinner, 1969, p. 47]. "To refuse,[8]" Skinner continues [p. 59], "is to leave further changes in our

[6]More explicitly, Dworkin [2020] defines paternalism as "interference of a state or an individual with another person against their will, and defended or motivated by the claim that the person interfered with will be better off or protected from harm". This is related to the concept of consent, especially consent of the governed, and to the distinction between positive liberty (the capacity to act upon one's free will) and negative liberty (freedom from external constraints) [Berlin, 1969], but a fuller discussion is outside the scope of this essay.

[7]See e.g. social contract theory, where citizens can be described as being "forced to be free" when they are constrained to obey the general will [Rousseau, 1762] or to obey a strong, undivided government [Hobbes, 1651].

[8]It would be unfair here to neglect to mention that in the same paragraph Skinner

culture to accident, and accident is the tyrant really to be feared."

3.3 Accident vs. Agency

We are torn, then, between the "fearful tyrant" of accident, and an absolute dictatorship.

The dictatorship of moral realism creeps into every area of life[9]. If we were purely maximising utility, for example, if would be intolerable to let a lone artist choose to paint a particular landscape, when an equally-accessible landscape nearby would make a more beautiful subject — we would be wasting an opportunity to create a superior artwork that could delight many.

Yet personal agency, self-expression, self-directed discovery: all these things are important to us, and seem a core part of the good life. A person's sense of control is associated with multiple physical and mental health outcomes (see e.g. the work of Keeton et al. [2008] on new parents, and of Rodin [1989] on older people). But health is merely a means to an end. Humans, while we may welcome certain constraints, feel a need to choose the bonds that bind us[10]. As Douglass [1855] said, of his bondage: "it was slavery — not its mere incidents — that I hated".

A positive definition of alignment, a single value to be optimised, is difficult to square with the desire to allow humans their own exploration, allowing them to make their own mistakes[11]. In the following section, I describe a possible such value.

---

is urging eugenics ("selective breeding"). It is interesting to note that this new social technology which in 1969 Skinner sees on the horizon lends an urgency analogous to that lent today by the advent of superintelligent AI. To quote from that same paragraph: "The "value judgements" which will then be demanded are beginning to attract attention."

[9]Though see Scheffler [1986] for a discussion about the limits of moral demands.

[10]Or at least some of them. Many proverbs encourage the acceptance of "things we cannot change", rather than futile frustration.

[11]It is interesting to note that exploration is core to many computational optimisation approaches, e.g. simulated annealing. Often a balance is struck between some kind of gradient-ascent, to find a local maximum, and random jumps to avoid getting stuck at a suboptimal local maximum. Eventually, the hope is that the best local maximum you find is the global maximum of the landscape, or at least a satisfactory solution. Accidents, perhaps, are inseparable from exploration of an unknown environment.

4 The Empowerment of others

In this section, I'd like to combine both the idea that I personally want the freedom to make choices that might contradict a benevolent dictator (i.e. that I want to make my own mistakes), and the idea that we may be forced to commit to a positive vision of AI alignment. I propose a unified target of AI alignment, which addresses the direct demand for some answer to the question of what to positively aim for, while avoiding overbearing paternalism, by prioritising the liberty of each individual — operationalised using the metric of Empowerment — and turning the question of a person's values back to them.

4.1 Sharing control

One solution to our dissatisfaction with the limitations of autocracy, and with our frustrations around external control, is the sharing and distribution of control, as seen in democratic ideals.

Here, the idea is that — roughly speaking — we all have similar strengths of desire and potential for pleasure and pain. We are all similarly likely to be correct or mistaken. Rules which can be agreed on are, we hope, desired by all participants. And so, we all decide to give each other similar amounts of power over each other. We all, in this ideal, take equal part in creating our moral and social environment (and, by extension, our physical environment and mental environments).

But these assumptions break down in a system where one entity is much more powerful than the others.

"Democracy is an effort to solve the problem by letting the people design the contingencies under which they are to live or — to put it another way — by insisting that the designer himself live under the contingencies he designs." [Skinner, 1969]

In the terminology of Skinner, it is hard to conceive of how a powerful designer could truly live inside the contingencies they design. Even if this designer were to create an avatar to exist inside the designed world, while the experience of the avatar might be comparable to that of most people, the experience of the designer would still be markedly different from those of the other inhabitants. They therefore would not be subject to the same conditions, and their incentives would be differently aligned.

To begin to get a handle on personal liberty in a way which is amenable

to being encoded in a machine-readable way, the following section describes the concept of Empowerment, used in information theory and robotics.

4.2 Overview of Empowerment

The concept of Empowerment in information theory formalises and quantifies the potential an agent perceives that it has to influence its environment [Salge et al., 2014]. It aims to correspond to the everyday notion of the word: if one picks up a key which opens a locked door, Empowerment should increase. It also tracks general and disparate drives such as "maintaining a good internal sugar level, staying healthy, becoming a leader in a gang, accumulating money, etc." [Salge and Polani, 2017], all acts which maintain and enhance one's ability to control the environment.

Empowerment is a measure of "At time t, where $t_A < t < t_S$, how much do I learn about which future states S_{t_S} are likely/possible, when I learn about which actions A_{t_A} were taken?"

Formally, it is defined as the 'channel capacity' between the set of future-environment-states S_{t_S} and the set of preceding actions A_{t_A}. That Empowerment is a rigorously defined mathematical object is important for it to be used as part of an objective function, and so a sketch of a formal definition is given in the appendix.

4.3 Subjectivity of Empowerment

The method of calculating Empowerment described above has a significant limitation: the representation of the State of the environment and the representation of Actions must exist as mathematical objects represented within the mind of the machine.

We are used to the idea of machines having internal representations of environments, and of them having internal representation of actions, and also with their conception of such states/actions potentially existing in the future. Coherent mathematical representations of set of states and actions exist, and so such metrics as Empowerment are calculable.

The difficulty arises when trying to understand the mind of another. Empowerment, crucially, depends on an entity's own *perceived* power / agency / liberty.

The exception to this is when one's perceived available-actions and predicted attendant future-environment-states is well-known by another. The

mind of a robot can be designed so that a programmer with complete access to the information it holds can also calculate its Empowerment.

But complete knowledge of another's mind is, in general, unattainable. How can a machine have an internal representation of an external mind?

4.4 Emulation to predict Empowerment

One approach would be to simulate the external mind. If a machine has an internal model of an external agent, then such an internal model can be fully known. Then, a metric of Empowerment would be available for the simulated entity, which could be used as a proxy for that of the external mind.

An agent then asking "How should I act?" could run an emulation of the system, to have visibility of the perceived environment-State and perceived potential-Actions of entities whose liberty it values, and calculate the Empowerment of these emulated beings in various scenarios. It could then, in real life, choose the action which had maximised the Empowerment of the emulated beings.

This might be thought of as a practical way to approach a form of preference utilitarianism, where the focus of the agent is on empowering the population to directly pursue their own preferences.

4.5 Limitations and challenges

This liberty-based form of a positive target of AI alignment consists of making space for human agency, even (or especially) where this may work against the agent's "better judgement". This comes across a number of challenges which must be addressed, some of which are familiar to political philosophy, though some details are significantly different.

4.5.1 Scope of beneficiaries of agency

The encoded preferences of a sufficiently powerful agent are likely to be near-maximally satisfied. If we assume that (despite the odds) the agent's creator successfully encodes their preferences in the objective function of the agent, then we may equally assume the flourishing of the agent's creator (assuming they retain control if their preferences change over time).

But what if we were to wish for a more distributed flourishing, perhaps encompassing all humans? (Or, if we take seriously the allegation of speciesism

popularised by Singer [1995], all sentient beings?)

We are also confronted by a question familiar to utilitarians: what distribution of liberty do we prioritise? Behaviour which maximises the average individual's Empowerment may differ from maximising total Empowerment, or from raising the level of the least-Empowered.

4.5.2 Meaningfulness of options

Can we weight actions or world-states by how meaningful they are, so that a machine does not remove a desired possibility in order to provide a multitude of petty choices? The meaningfulness of an outcome can be difficult to distinguish: if one misses a train, it is largely irrelevant that many other trains to different destinations are now available to you. Sufficiently powerful prediction of outcomes goes some way to addressing this, but an answer is needed for instances where an agent is unsure of the outcomes of its behaviour.

While these questions must be answered for a liberty-based positive definition of the target of AI alignment to be useful, none of these questions seem unanswerable.

5 Conclusion

AI alignment is a necessary prerequisite for a recognisably good life after the creation of an arbitrarily intelligent and powerful entity, which is a potential outcome of superintelligence. Experts expect the creation of superintelligent AI within this century.

This essay has argued that merely negating a collection of negative definitions — definitions of outcomes we wish to avoid — is not sufficient to ensure AI alignment. The essential nature of optimisation at the core of most AI implementations is to find unusual and extreme outcomes and behaviours and, historically, many of these have not been foreseen. A superintelligent agent is likely to resist value update, so we may not be able to rely on iteratively correcting our instructions as we realise their shortcomings.

A positive target for AI alignment must take one of two forms. In the first, a direct definition of human flourishing is decided upon, such as happiness, and a metric is created; the agent then works to maximise this metric. This essay's main dissatisfaction with this direct approach is that human agency is then squashed, and humans find their agency to be important even when —

as in the case for the slave who is generally left alone — the "mere incidents" of their existence may otherwise be identical.

The second form of a positive target of AI alignment consists of making space for human liberty and agency, even (or especially) where this may work against the agent's "better judgement". This avoids the problems of the first form, while still functionally taking the practical form of metric maximisation.

This second form comes across a number of challenges which must be addressed, some of which are familiar to political philosophy, though some details are significantly different. While these questions must all be answered for a liberty-based positive definition of the target of AI alignment to be useful, none of these questions seem unanswerable. I therefore conclude that a liberty-based positive definition of AI alignment's target is possible.

This essay also considers the question of how to operationalise a quantity such as personal liberty: the information-theoretical metric of Empowerment is used here as an example of the kind of mathematical object which is needed. Since Empowerment is a naturally subjective measure, emulation of other minds is proposed as a solution: a machine could measure the Empowerment of minds which it simulates, and act in ways which would be best in the emulation. The problem then becomes a practical one of prediction.

One speculative outcome of such a situation comes from combining it with the potential that our own existence is within a computer simulation [Bostrom, 2003], a possibility which raises the question of the intentions of the simulator. A liberty-aligned agent who was trying to determine what behaviour it *should* engage in might simulate minds, to ask implicitly of them the question "what do you want?".

Perhaps, then, the point of life is to enjoy it.

Appendix:
Sketch of a formal definition of Empowerment

A sketch of a formal definition of Empowerment is given here. A more detailed introduction is given in [Salge et al., 2014].

Entropy (the uncertainty about a discrete random variable X before observing it) is defined as:

$$H(X) = -\sum_{x \in X} p(x) \log p(x), \tag{1}$$

where x is a value taken from X with probability $p(x)$. If another variable $y \in Y$ is introduced, we can define the conditional entropy of "X given Y" (the remaining uncertainty about X when Y is known) as:

$$H(X|Y) = -\sum_{x \in X} p(y) \sum_{y \in Y} p(x|y) \log p(x|y). \tag{2}$$

We can now define Mutual Information, which is the average amount of information about X which is gained by observing Y (or vice-versa, since Mutual Information is symmetric):

$$I(X;Y) = H(Y) - H(Y|X). \tag{3}$$

The Empowerment E is then defined as the channel capacity C between A and S, which is the maximum mutual information between states and actions. If we restrict to two moments in time, t and $t+1$, then:

$$E = C(A_t \rightarrow S_{t+1}) = \max_{p(a)} I(S_{t+1}; A_t) \tag{4}$$

where the maximum is taken over all possible choices of a (where $p(a) > 0$).

References

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

Appel, K. and Haken, W. (1989). *Every Planar Map is Four Colorable*. American Mathematical Society.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula.

Begon, J. (2016). Paternalism. *Analysis*, 76(3):355–373.

Berlin, I. (1969). Two Concepts of Liberty. In *Four Essays on Liberty*.

Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly*, 53(211):243–255.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 1st edition.

Bourget, D. and Chalmers, D. J. (2013). What do philosophers believe? *Philosophical Studies*, 170(3):465–500.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Douglass, F. (1855). *My Bondage and My Freedom*.

Dworkin, G. (2020). Paternalism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2016). The off-switch game.

Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Knopf Doubleday Publishing Group.

Hales, T. (2005). A proof of the Kepler conjecture. *Annals of Mathematics*, 162(3):1065–1185.

Hobbes, T. (1651). *Leviathan*.

Keeton, C. P., Perry-Jenkins, M., and Sayer, A. G. (2008). Sense of control predicts depressive and anxious symptoms across the transition to parenthood. *Journal of Family Psychology*, 22(2):212–221.

Krahulik, M. and Holkins, J. (2004). Gabriel the Orator.

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

Mill, J. (1974). *On Liberty*.

Müller, V. C. and Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence*, pages 555–572. Springer International Publishing.

Rawls, J. (1993). *Political liberalism*. Columbia University Press, New York.

Rodin, J. (1989). Sense of control: Potentials for intervention. *The Annals of the American Academy of Political and Social Science*, 503(1):29–42.

Rousseau, J.-J. (1762). *The Social Contract*.

Russell, S. (2019). *Human compatible : Artificial Intelligence and the problem of control*. Allen Lane/Penguin Books, London.

Salge, C., Glackin, C., and Polani, D. (2014). Empowerment – An Introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer Berlin Heidelberg.

Salge, C. and Polani, D. (2017). Empowerment as Replacement for the Three Laws of Robotics. *Frontiers in Robotics and AI*, 4.

Scheffler, S. (1986). Morality's Demands and Their Limits. *The Journal of Philosophy*, 83(10):531.

Schmid, M., Moravcik, M., Burch, N., Kadlec, R., Davidson, J., Waugh, K., Bard, N., Timbers, F., Lanctot, M., Holland, Z., Davoodi, E., Christianson, A., and Bowling, M. (2021). Player of games.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359.

Sims, K. (1994). Evolving virtual creatures. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques - SIGGRAPH '94*. ACM Press.

Singer, P. (1995). *Animal Liberation*. Random House.

Skinner, B. F. (1969). *Contingencies of reinforcement : a theoretical analysis*. Appleton-Century-Crofts, New York.