Friendly superintelligent AI: All you need is love

Michael Prinzing

University of North Carolina, Chapel Hill, North Carolina 27514, USA

Abstract. There is a non-trivial chance that sometime in the (perhaps somewhat distant) future, someone will build an artificial general intelligence that will surpass human-level cognitive proficiency and go on to become "superintelligent", vastly outperforming humans. The advent of superintelligent AI has great potential, for good or ill. It is therefore imperative that we find a way to ensure—*long before* one arrives—that any superintelligence we build will consistently act in ways congenial to our interests. This is a very difficult challenge in part because most of the final goals we could give an AI admit of so-called "perverse instantiations". I propose a novel solution to this puzzle: instruct the AI to love humanity. The proposal is compared with Yudkowsky's Coherent Extrapolated Volition, and Bostrom's Moral Modeling proposals.

1 Introduction

Many AI researchers believe there is a non-trivial chance that AI with greater than human-level cognitive capabilities will be developed sometime in the (perhaps somewhat distant) future (Müller & Bostrum 2016). These AI may eventually become "superintelligent"—i.e., capable of *vastly* outperforming humans in all, or nearly all, cognitive tasks (Bostrom 2014; Chalmers 2010).[1] This is both an exciting and unsettling prospect. If a superintelligence were friendly—if its goals aligned with ours—it could provide untold benefits to humanity. It could cure diseases. It could increase economic output such as to end all poverty, hunger, and need. On the other hand, if the superintelligence were malignant—if its goals diverged in important ways from ours—we would be powerless to stop it (Bostrom 2014, chapters 8-10).[2]

There is always a risk, when thinking about topics like this, of falling into futile futurology: wild, ungrounded speculation. But, while it's good to keep this risk in mind, I believe it is important to start considering the possibilities. For one thing, the suggestion that superintelligent AI will (eventually) be developed is *not ungrounded*. This is something that many AI researchers are seriously concerned about (Müller & Bostrum 2016; see also Bostrom 2014; Soares & Fallenstein 2015; Yampolskiy 2016). This certainly doesn't mean that superintelligence is inevitable—just that it's a real possibility. Thus, given the magnitude of the possible outcomes, it is important to work out

[1] There are a number of ways in which an "intelligence explosion" like this might happen (Good 1965). Perhaps the most plausible would involve a seed AI undergoing recursive self-improvement (Yampolskiy 2016, chapter 5).

[2] Given that superintelligence is defined as cognitive performance *vastly beyond human-level*, this is hard to dispute.

how we can increase the probability of ending up with a friendly superintelligence, and decrease the probability of a malignant one. It's far better to have a solution and not need one, than to need one and not have one. Crucially, this work will need to be done before (preferably long before) any such being is created. The time to start on this problem is *now*.

It seems to me that the way to understand the friendly superintelligence problem is in terms of a relationship. What we are aiming to ensure is that our relationship with the superintelligence is healthy and beneficial. Perhaps the only such relationship we can have with such a being is no relationship at all. But, in any case, we are trying to ensure that it wants what is best for us, cares about our interests, and will do right by us. When the question is put it in these terms, the shape of a solution starts to become clear. What will make our relationship with the superintelligence a good one is the same thing that makes *any* personal relationship good: love.

2 The puzzle: Avoiding "perverse instantiations"

Before I explain this proposal, it's important to specify the puzzle it aims to solve. There are several related puzzles that it is *not* a solution to. These concern how to express complex, world-affecting goals in computer code. Suppose the goal we want the AI to pursue is the maximization of human happiness. It's quite easy to state such a goal. But it's extremely difficult, even in a natural language like English, to explain precisely what it means. This is classic philosophical territory. *What is happiness?* Supposing that we could give a correct accounting of the relevant concepts, we would then face the challenge of converting that natural language analysis into the functions and operators of a programming language. These initial puzzles, in short, are concept learning puzzles. Abstract natural language concepts are the ingredients of any goal that we might give an AI. The issue I'm interested in, for the purposes of this paper, assumes that we have solved these puzzles.

The next challenge, the one I'm addressing, is to figure out which goals or instructions to provide. A direct approach would be to simply specify some concrete goal like "maximize human happiness". The main problem with this method, as Nick Bostrom (2014) discusses, is that it is extremely difficult to accurately capture the things we value in explicit specifications. For any goal that we might specify directly, even one that sounds agreeable, there is a significant chance that realizing the goal would go horribly wrong. In Bostrom's (2014, 146) terminology, that goal likely admits of some "perverse instantiation"—a way of realizing the letter of the instructions, while betraying their spirit. For instance, one way of maximizing human happiness would involve implanting electrodes into our brains and inducing a perpetual state of vacuous, drooling euphoria. We *would* feel happy, but that's not exactly my vision of a life well-lived.

Thus, those who have thought seriously about this problem tend to favor a more indirect approach (Yudkowsky 2004; Bostrom 2014). Rather than specify a concrete goal, we should give the AI a procedure by which to decide for itself what to do. This

approach is appealing insofar as it delegates to the superintelligence the hard work of figuring out how to remain human-friendly. My proposal adopts this indirect strategy.

3 The solution: Love

I propose that we can avoid perverse instantiations by instructing the AI to love humanity. We could start with a roughly human-level AI—one which we expect to undergo recursive self-improvement resulting in superintelligence—and give it the final goal of loving humanity. As it progresses towards superintelligence, the AI's understanding of love (its grip on the concept and its referent) will deepen. Since the AI's final goal is to love humanity, it will seek to better understand both what love is (what we have in mind, as well as what we refer to when we use the term in the relevant ways), and what is involved in being loving (if Abe loves Bea, what implications does this have for his behavior?). Ultimately, its goal would be to apply what it has learned about love to its relationship with us.[3] The following sub-sections elaborate this proposal.

3.1 What is love?

One might worry that the concept of love is simply to slippery and too vague to be of use here. However, it's possible to provide a good deal of precision to the idea. As psychologist Beverly Fehr (2013, 202) writes, there "are now empirically based answers to questions about the meaning of *love* and how it is experienced". Obviously, "love" has many senses: "I love pizza"; "I am in love with her"; "I love you man". This is because the term is not applied to a single phenomenon, but a web of related phenomena.[4] Most importantly for our purposes, empirical investigation has distinguished four kinds of love (Berscheid & Hatfield 1974; Berscheid 2006; 2010; Fehr 2013): romantic/passionate, attachment, companionate, and compassionate. The relevant forms for my proposal are the final two.

It's worth emphasizing what I'm *not* suggesting. Romantic or passionate love—what the ancient Greeks called *eros,* and which English speakers pick out with the expression "being *in* love"—is typically triggered by physical attraction and sexual desire. Naturally, this is not what I have in mind. Nor am I suggesting that the superintelligence display attachment love, affection directed towards a particular individual. What I'm proposing is that the superintelligence display companionate and compassionate love towards humanity. Companionate love is the kind found between close friends and family members. It's characterized by caring, trust, honesty, respect, and can be experienced for many people simultaneously. Compassionate love is sometimes

[3] Some reviewers have questioned whether AI will be capable of emotion. I see no grounds for skepticism, however. Artificial emotions have long been a theme in AI research (Sloman & Croucher 1981; Picard 1997; Scheutz 2014).

[4] "Perhaps it is no wonder that love has puzzled so many for so long. Part of the confusion is that the word 'love' has been affixed to different parts of this larger, dynamic love system" (Fredrickson 2016, 848).

called altruism or selfless love. "A unique antecedent of compassionate love is the perception that the other is in distress or in need" (Fehr 2013, 203). I'll say more about these forms of love when I discuss what we should expect from a loving superintelligence in Section 3.3. The point, for now, is that love is not a hopelessly vague notion. There are precise, empirically grounded ways to articulate the idea.

3.2 Who is to be loved?

Thus far, I've claimed that the superintelligence should love humanity. But who *exactly* should this include? Should it include fetuses, and brain dead comatose patients? Or, perhaps "humanity" is too narrow. Perhaps, we should include non-human animals and extra-terrestrial life (if there is any). I'm inclined to say that the superintelligence should love all persons. This specification brings us into philosophically treacherous waters. Who counts as a person is itself disputed. We might choose to err on the side of inclusiveness, allowing in all disputable categories just to be safe. Then again, this might beg important moral questions in favor of vegans and pro-lifers. A superintelligence that loved fetuses and non-human animals might prevent abortions and meat eating, which could be a wrongful imposition. Perhaps the best option will be to defer to the superior epistemic position of the superintelligence by instructing it to discover the extension of the term "person". This would add greater complexity to my proposal. But, it may be worth it.

3.3 What should we expect from a loving superintelligence?

While it is impossible to say in detail what a loving superintelligence would do (if we could say, then we would be superintelligent ourselves), the psychology and philosophy of love provide strong evidence for some general predictions.

There is robust psychological evidence that companionate love is the most central part of the concept of love. When subjects in psychological studies are asked about the characteristics of love, they tend overwhelmingly to cite features of companionate love (Fehr 1988). These results have been replicated repeatedly and display cross-cultural stability (Luby & Aron 1990; Button & Collier 1991; Kline et al. 2008). Moreover, in a series of studies, Fehr and Russell (1991) found that companionate love was considered the most typical or paradigmatic form. Five features of love are consistently (across studies and cultures) found to be central to the concept: trust, caring, honesty, friendship, and respect (Fehr 2013, 206).

Somewhat recently, evidence has emerged that the concept of love may be essentialist—i.e., "people view certain features as *necessary* for them to judge that a given relationship is an instance of the concept of 'love'" (Hegi & Bergner 2010, 634). The necessary feature is "investment in the well-being of the other, for his or her own sake" (Hegi & Bergner 2010, 621). When investment in the other's interests was described as missing from a relationship, subjects consistently found it "very contradictory" to assert that the people loved each another. In other words, if Abe is not invested in Bea's well-being for her own sake, then it simply can't be true that Abe loves her. Though other characteristics of love are important, none display this *sine qua non* status.

In other words, the psychological evidence strongly supports the claim that "love, by definition, conveys a caring orientation toward others" (Fredrickson 2016, 850). An essential part of a father's love for his daughter, for instance, is his desire to see her flourish. Depending on his beliefs about what that requires, this may mean taking an active role in her development or stepping back to allow her to struggle and grow on her own. We would see something analogous from a loving superintelligence.

This conclusion is reinforced by consideration of the philosophical literature (Helm 2017). Two of the three standard philosophical theories of love—love as robust concern (Frankfurt 1999), and love as emotion (Badhwar 2003)—concur with these psychological conclusions.[5] The love as emotion view effectively defers the question "what is love?" to psychology. And, interestingly, the love as robust concern view just is the conclusion drawn from Hegi & Bergner 2010. Love, on the robust concern view, is a robust concern *for the beloved's well-being*. To love someone is to take on her interests as your own.

Thus, the psychology and philosophy of love give us strong reason to expect that a superintelligence which loves humanity would be invested in our well-being for our own sakes. It would display trust, caring, honesty, friendship, and respect. By giving the AI the final goal of loving humanity, we not only rule out nightmarish scenarios in which the superintelligence completely disregards human interests, we also ensure that it seeks to *advance* our interests.[6]

3.4 What happens when interests conflict?

It's important to recognize that loving someone doesn't always mean giving her what she wants (or thinks she wants). Suppose that Bea would like a house by the sea. If Abe loves Bea, and if he sees no reason *not* to give Bea a house by the sea, then he will try to give her one. But there will be many cases in which Abe's love for Bea does not lead him to do this—for instance, when Abe thinks that, despite what Bea believes, a house by the sea is not in her interests. Thus, while a loving superintelligence would be invested in human interests, it would not attempt to uncritically satisfy each human desire. A particularly important instance of this phenomenon arises from interpersonal conflicts.

Sadly, some humans are hateful. They have enemies with whom they fight, and whom they wish they could destroy. Thus, one might object, if a superintelligence were to love Ayman al-Zawahiri (the leader of al-Qaeda), would it not seek to destroy his

[5] The third theory, love as valuing (Singer 2009), likely concurs as well. Though explaining how would take us too far afield.

[6] Of course, in order for it to do this, it must have sensible views about what our well-being consists in. A religious zealot might sincerely think that she advances my well-being by forcibly converting me. It seems pretty clear, however, that the zealot's beliefs about well-being are false. (Were it not for her belief in an afterlife, she would probably reject them herself.) There are currently vibrant research programs in psychology and philosophy, which have revealed a good deal about the nature of well-being (for surveys see Snyder & Lopez 2009; Fletcher 2016). A superintelligence, with its superior epistemic position, can be expected to have even better-informed views about well-being.

enemies, and to advance his wicked goals and ambitions? Don't those who love evil people do evil things for them?

This objection overlooks *why* some humans do evil things for the people that they love. The fact is that humans don't love everyone. And they certainly don't love everyone equally. One of the ways in which love-based reasons for action can be outweighed or undercut is by their conflict with *other* love-based reasons. Imagine that two of your equally beloved friends are fighting. Your friend Abe becomes enraged by a dispute with your friend Bea. He wants nothing more than to see Bea forlorn and destitute. Given that you also love Bea, is there any chance that you will satisfy Abe's desire? I think not. In this case, your love for Bea undercuts your reason to give Abe what he wants. Since my proposal is that the superintelligence be programmed to love everyone equally, it would not harm one human merely to satisfy another's malevolent desire.

If Abe and Bea are so belligerent that they each demand you take a side, the most loving thing to do, I suspect, would be to withdraw from both. If they each want nothing to do with you unless you take sides and hate the other, then the only remaining choice is to distance yourself from both and hope that one day a resolution can be found. A love egalitarian superintelligence would likely display similar behavior. Of course, a relationship with a loving superintelligence would be *extremely* advantageous. As I've indicated, such a being could potentially do wonderful things for us. Thus, there would be very powerful incentives to maintain a close relationship with the superintelligence. This would require one to not be at odds—or, at least, to not be too aggressively adversarial—with the other objects of the superintelligence's love (i.e., other people). This may be an added benefit of my proposal. A loving superintelligence could potentially help to resolve some of humanity's conflicts as well as prevent new ones.

4 Comparison with Alternatives

My proposal has two main competitors: Eliezer Yudkowsky's (2004) Coherent Extrapolated Volition, and Nick Bostrom's (2014) Moral Modeling. I'll compare my proposal to each.

4.1 Coherent Extrapolated Volition

Yudkowsky has proposed that superintelligent AI should promote our "coherent extrapolated volition" (CEV). This proposal takes what philosophers would call an "ideal advisor" approach, one that centers on what we would desire under idealized conditions. Yudkowsky writes:

> [O]ur coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted. (Yudkowsky 2004, 6)

Since this proposal is put in intentionally poetic terms, some elaboration is necessary.

To "know more" means to be aware of all the relevant facts concerning the objects of our decision-making. If we "thought faster" we would not just be smarter; we would also have thought longer and more clearly about our desires and options in life. Some of our personal characteristics and desires receive second-order endorsement; some don't. For instance, Abe may approve of his affection for his friends, while wishing he were less angry and vindictive. If he "were more the person he wished to be", he would keep or increase the former trait, while reducing or eliminating the latter. If we "had grown up farther together" we would have had more shared experiences and a stronger sense of solidarity with humanity. Extrapolation "convergence" refers to a high probability of one's choosing in a certain way under idealization. If it's hard to know what our idealized selves would choose concerning something, then the superintelligence is to leave those options open. The points on which individual extrapolated volitions "cohere rather than interfere" are the choices that idealized humans would agree on. The idea is to have the superintelligence only act on those goals that everyone (or maybe just most people) would endorse in their idealized state. The requirement that our volition be "extrapolated as we wish that extrapolated, interpreted as we wish that interpreted" indicates that our present selves should guide the process of idealization. Thus, the resulting extrapolated volition should include goals and/or values that our *actual* selves could be led to understand and approve of.

The idea, in short, is that the superintelligence is to take each individual, idealize her belief and desire set in order to determine a "volition", and then aggregate these volitions, acting where they cohere. I believe that my proposal captures what is appealing about this proposal, while avoiding some of its problems. I'll discuss each in turn.

Motivations For CEV. The central appeal of this proposal is that it leaves humanity in control, in some sense, of our future. The superintelligence would do whatever we would have collectively and ideally decided for it to. As I've indicated, a loving superintelligence will, in a similar fashion, seek to advance human interests, but will not blindly satisfy expressed desires. It will aim to do what is *best* for us.

Yudkowsky (2004, 14) suggests that another appeal of his proposal is that it allows for moral progress. There are some living today who remember racial segregation in the United States. A mere 200 years ago chattel slavery was alive and well. It seems unlikely that moral progress has peaked at this moment in history. In a few decades or centuries, our own moral sensibilities may look as barbarous as segregation and slavery do now. The CEV proposal aims to accommodate this thought. I believe love egalitarianism would do the same. For now, however, readers will have to accept a promissory note: I'll say more about the morality of a loving superintelligence in Section 4.2.

Yudkowsky's proposal would have the superintelligence only act where our extrapolated volitions cohere. This is thought to prevent a particular subset of humanity form "hijacking" our planet's (or universe's) future, imposing its own conception of the good on all of humanity forever (Yudkowsky 2004, 17). Imagine, for instance, how you would feel if it were al-Qaeda that first built a superintelligent AI and programmed it to advance their goals. By ensuring that the superintelligence advances goals that we all can agree on, Yudkowsky thinks, we can avoid potential conflicts over how the AI

is to be programmed. (Imagine what the Pentagon might do if it became convinced that al-Qaeda were about to develop superintelligent AI.) This will be very important because an AI arms race increases the incentive to cut safety corners, increasing the risks of malignant superintelligence (Armstrong et al. 2013).

A loving superintelligence would similarly prevent any one person or group from hijacking humanity's future. As I emphasized in Section 3.4, the AI is to be programmed to love everyone equally. Since love involves investment in the well-being of the beloved, and since autonomy is an important constituent of human well-being (Ryan & Deci 2000; Helm 2017, Section 3), a love egalitarian superintelligence would be strongly averse to imposing foreign conceptions of the good on us.

Problems with CEV. There are two problems with the CEV proposal: the idealization of our actual selves adds enormous complexity; and there will likely be insufficient coherence in the individual extrapolated volitions.

The computational resources required to extrapolate a person's volition would be astronomical. In rough sketch, discovering Abe's extrapolated volition would require the AI to: either observe Abe's behavior or (for better accuracy) scan his brain; discover and categorize his present beliefs, desires, and interests; search for inconsistencies; discover all the potential objects of his volition; model his learning and thinking processes, and then amp them up so that he "thinks faster"; model various social interactions that Abe might experience which would lead to a wiser, more mature and pro-social version of Abe; filter those out from other social experiences that would make him less wise or pro-social; search again for inconsistencies that arise as his desires are extrapolated; resolve any such inconsistences; produce a rough draft of his volition and compare it with the actual Abe to see whether real Abe would approve of his idealized self and his idealized choices. This would be insanely computationally demanding. If we then multiply this by 8 billion or more people, it becomes unfathomable. After all the individual volitions have been extrapolated, there also remains the task of synthesizing them for coherence.

Perhaps my imagination is simply too limited. But I have a hard time imagining even a superintelligence with the processing power necessary to complete the task. Perhaps the CEV could be computed if all of Earth's resources were converted into processing power. But, of course, by then the CEV is irrelevant. It will be too late to do us any good. On the other hand, maybe the superintelligence needn't actually determine humanity's CEV precisely. Bostrom (2014, 213) suggests that a superintelligence would be able to make a pretty good *guess* as to what our CEV would include, and then act on its guess. It could then update its speculation as it learns more about human psychology and society.

> For example, it is more plausible that our CEV would wish for there to be people in the future who live rich and happy lives than that it would wish that we should all sit on stools in a dark room experiencing pain. If *we* can make at least some such judgments sensibly, so can a superintelligence. (Bostrom 2014, 213)

This seems right. But—given how much is at stake—I'd want a hell of a lot more precision than that! Part of what makes this kind of guesswork so difficult is the absence

of any examples. Present-day machine learning works best with large data sets—lots of example cases to train algorithms. But, we have zero examples of extrapolated volitions. Perhaps future technologies won't require exemplar data. But, we have no way of anticipating such developments.

Of course, some of the complexity required by the CEV proposal would also be required by my proposal. In either case, for instance, the AI will need to know a lot about individual human desires. But, my point is that by basing the AI's instructions on an *idealized* version of each individual, the CEV proposal is vastly more complicated than mine. On my approach, the superintelligence would have much needed examples. It could see actual loving relationships and use those as (fallible) models for its own relationship with humanity. Just like discovering our CEV, discovering what love involves would also require studying human psychology and society. But it won't require the additional (and, I think, Herculean) task of idealizing that psychology and society in order to determine what we would collectively wish for in some distant counterfactual scenario.

Another problem for the CEV proposal arises from the fact that it would have the superintelligence only act where our extrapolated volitions cohere. This, as we saw, was taken to be one of its merits. But, there is likely to be little coherence. It's certainly true that most humans desire things like food and shelter, opportunities to socialize and express themselves, and so on. These would plausibly end up in each extrapolated volition. But, beyond basic needs like these, there is likely to be little convergence on a wide range of important issues. Consider all the different kinds of people that there are: acetic hermits and minimalists, and materialistic Wall Street money-chasers; religious fundamentalists, and anti-theists; neo-Nazis, and social justice warriors; right libertarians, and communists. Even if all these people were idealized in the relevant ways, it seems highly unlikely that there will be much coherence in their volitions on questions of politics, economics, morality, and so on. If that's right, then, the superintelligence's hands would be tied on these important issues.

A loving superintelligence will similarly find itself in a position where the various objects of its love have incompatible interests. However, this will not prevent it from acting on those interests. One obvious way to satisfy conflicting interests is to localize them. If the hedge fund manager wants to live in a post-industrial, free market society, and the acetic hermit wants to live in an agrarian, barter-based society, then they could each have their own corner of the planet (or, if we're advanced enough, a planet of their own) on which to do that. Now, you might think that the *idealized* hermit and hedge fund manager would also reach this compromise, so this isn't an advantage for my proposal. I'll concede that this is possible. However, as we have no examples of extrapolated volitions, we simply can't be sure that they would. Moreover, even if the extrapolated volitions of the hermit and hedge fund manager end up with a high degree of coherence, it's far from obvious that this would generalize. It is an unfortunate fact that many humans' desires are not merely incompatible with the desires of others—but, worse, precisely what they desire is the thwarting of others' desires. (Recall the al-Zawahiri example from above.) Some of these nasty desires might be eliminated by the volition extrapolation process. But, given that one's extrapolated volition is meant to be sensitive to one's current self, it's hard to imagine that, for instance, the neo-Nazi's

extrapolated volition would bear much good will towards the social justice warrior. Similarly, it's hard to imagine the social justice warrior's volition cohering—or even compromising with—the neo-Nazi's. Thus, if we instruct the superintelligence to advance only those goals that we would all agree on, then we give up on much of the good that a superintelligence could do for us.

4.2 Moral Modeling

Bostrom mentions an additional concern regarding the CEV proposal. This is that even our idealized selves might not be such great people. "Moral goodness", he writes (Bostrom 2014, 267), "might be more like a precious metal than an abundant element in human nature, and even after the ore has been processed and refined in accordance with the prescriptions of the CEV proposal, who knows whether the principal outcome will be shining virtue, indifferent slag, or toxic sludge?"

This consideration suggests another approach, which Bostrom calls Moral Modeling (MM). The idea is to program the AI to learn moral concepts, discover moral facts, and promote moral goodness and/or rightness. If the superintelligence is better at moral philosophy than humans are, then it may succeed where we have yet to—discovering and acting on the correct moral theory. In other words, we could make a super-moral superintelligence.

Assuming that this approach could be successfully implemented, it would surely result in the *morally* best results of all the options. There are, nevertheless, a few reasons why I prefer my proposal. First, I suspect that the behavior of a love egalitarian superintelligence and a super-moral superintelligence would be very similar. If there are conflicts between morality and love egalitarianism, it's not clear to me where they lie. Obviously, my inability to spot differences does not entail that there are none. The thought is just that the expected outcomes for MM and my proposal will be quite similar, perhaps even identical.[7]

Some philosophers have seen a tension between love and morality—in some cases treating them as distinct, even competing, domains within practical reasoning (Slote 1983; Wolf 1992). Morality, they suggest, is supposed to be impartial; it's about taking into account everyone's interests, and not weighting some people as more important than others. Love, on the other hand, is inherently partial. If I love my friend, I will

[7] It has been suggested to me that love egalitarianism might be operationally equivalent to an interest-based consequentialism. If loving someone means being invested in her well-being, then loving equally should require weighing and advancing individual interests equally. Something like this is probably right. But it also seems clear that love comes with deontological constraints. If I killed a healthy person in order to harvest his organs and save five other people, I could not plausibly insist that I nevertheless loved him. This shows why we couldn't simply instruct the AI to promote human well-being. "Promote" is vague. Does it mean: maximize the sum total? maximize the minimum individual level? maximize the maximum level? satisfice to some threshold?… Love helps to resolve this problem. Loving is not a maximizing procedure, and (as I suggested) comes with side constraints. Though we can't articulate the procedure for making loving decisions, we clearly follow some such procedure in our daily lives. And we seem to think it's the right way to do things.

favor him over others. Thus, love and morality can, and perhaps often do, conflict. This conflict is dissolved, however, if we recall that the superintelligence will be programmed to love everyone equally. What is impartiality, after all, if not equal partiality towards all?

It's also worth mentioning that some moral philosophers think that even the appearance of tension between love and morality is illusory. As David Velleman (1999, 341) writes, "Love is a *moral* emotion precisely in the sense that its spirit is closely akin to that of morality. The question, then, is not whether two divergent perspectives can be accommodated but rather how these two perspectives converge".[8] In the tradition of Christian ethical thought, love plays a very central role. Paul Ramsey, a major figure in twentieth century Christian ethics, argued that love is the basis for all of moral theory. Ramsey (1950, xvi-xvii) argued that love is the "primitive idea" and "fundamental notion" for morality. Arthur Dyck (1968), another theologian, argues that love is not only a moral virtue, but the primary—perhaps sole—test and guide for action. As he puts it, "love is no mere sentiment or emotion. It is the relational bond of a covenant to form and sustain community… But it is also the power and the passion to get on with this task" (Dyck 1968, 545). Some Christian ethicists argue that a complete moral theory will have to incorporate other concepts or principles that cannot be derived from love alone (Harris 1976). But, regardless, it's clear that love has a central place in this tradition of ethical thought.

In short, I expect that a superintelligence programed to love to humanity would get us most—if not all—of what we would hope for in a super-moral superintelligence. So, in terms of outcomes, the choice between these proposals may not make much of a difference. In terms of difficulty in implementation, and the probabilities of success, however, my proposal has an advantage. Our goal, recall, is to maximize the probability of good outcomes and minimize the probability of bad outcomes. I believe that my proposal is superior because it comes with less risk of something going wrong. There are several reasons why that would be.

Working out what would be necessary in order to implement MM reveals layers of added complexity. Before we could give an AI instructions for discovering moral facts, we would need to figure out what moral facts are and how they can be discovered. In other words, we need to answer the central questions of metaethics (moral semantics, metaphysics, and epistemology). This would be an enormous initial hurdle. A natural thought here would be to outsource this work to the AI itself. Before it does any first-order moral theorizing, the superintelligence should work out the correct metaethical theory. So, our instructions might be: "Figure out what moral facts are, and how they might be discovered. If there are any, discover what they are. If there are no moral facts, or if moral facts are culturally relative, or some such thing, then shut down. Oth-

[8] On Velleman's view, respect for others is the minimum of moral expectation, while love is the maximum of moral supererogation. "[R]espect is a mode of valuation that the very capacity for valuation must pay to instances of itself. My view is that love is a mode of valuation that this capacity *may* also pay to instances of itself. I regard respect and love as the required minimum and optional maximum responses to one and the same value" (Velleman 1999, 366).

erwise, perform the morally best actions." I'm a pessimistic about this approach, as I'll explain.

When it comes to morality, people disagree a lot. Philosophers and non-philosophers alike disagree about which moral concepts apply to which objects of evaluation (e.g., which actions are right, which character traits are vicious). They also disagree about what makes it the case that moral concepts apply. Many people believe that God's will is what makes an action right or wrong. Others think that God's will, should it exist, has no important connection with the right- or wrong-making properties of an action. Some people think that moral facts are culturally relative. Others think that they are absolute and response-independent. Some think that moral evaluations are evaluations of outcomes or states of affairs. Others think that moral evaluations are evaluations of a person's will, intentions, or character. In other words, there is substantial disagreement at every level of moral discourse (Merli 2009). Perhaps no other concepts are as disputed as moral concepts. Some even claim that they are "essentially contested"—meaning that this kind of disagreement is an essential or constitutive feature of the concepts (Gallie 1955).

All of this conceptual disagreement would make it extremely difficult for an AI to make sense of moral concepts. I'm not taking a stand here on the metaethical implications of moral disagreement. My point is that, plausibly, an AI's only access to moral concepts (and thereby moral facts) will be through human moral discourse. I'm assuming that a machine would not have independent access to moral reality. On some metaethical views, an AI could have direct epistemic access to mind-independent moral facts through the capacity for reason, or a faculty of moral intuition, or some such thing. However, even on the assumption that such a view is right, it's not at all clear what such faculties are or how they would work—much less whether and how they could be incorporated into an artificial being. It is far more plausible that, if a machine is to discover moral facts, it will have to do so through us. Thus, widespread and persistent conceptual controversies, which would make it very difficult for an AI to acquire the concepts, pose a serious obstacle to a successful implementation. The motivation for MM was to prevent human foibles and moral imperfections from spoiling the good that a superintelligence could do. But, given that it will have to acquire its moral knowledge though us, it seems that this proposal doesn't actually solve that problem.[9]

[9] Of course, people sometimes also disagree about what's involved in loving someone. But, this is not typically disagreement about what love *is*. As the research surveyed in Section 3.3 showed, there is a remarkable degree of consensus on that question (despite appearances). One might object, in a similar spirit, that there is disagreement about what well-being consists in. Since loving involves an investment in well-being, if well-being is as controversial as morality, then my view has the same problem as MM. I deny, however, the antecedent of this conditional. While there certainly are competing theories of well-being, for the most part, there isn't much disagreement in the literature over what contributes to a person's well-being (see Fletcher 2016). People tend to agree on which things are good for a person, even if they disagree about *why* those things are good for her. (E.g., is accomplishment intrinsically good for a person, or only insofar as it contributes to his positive mental states?) When it comes to the practical matter of promoting well-being, however, such disputes may not be of much significance.

5 Conclusion

It is vitally important that we figure out, before superintelligence is developed, how to ensure that it acts in ways congenial to human interests. I have suggested that we think of this problem in terms of a relationship. And the key to a good relationship is love. Thus, my proposed solution to the problem of friendly superintelligence is to teach the AI about companionate and compassionate love, and instruct it to love everyone equally.

After briefly clarifying this proposal, and exploring some of its implications, I compared it with two of the most promising alternatives: CEV and MM. I argued that my proposal captures what is appealing about the CEV and avoids its most serious problems. I also argued that the outcomes of my proposal would likely be very similar to the outcomes from a successful implementation of the MM proposal. However, implementing MM would face greater challenges. Thus, my proposal is to be preferred.

This paper is intended to open a line of inquiry. Obviously, I don't pretend to have resolved the problem of friendly superintelligence. Rather, I've suggested—at a very general, non-technical level—an approach for solving the problem. If the ideas presented here hold up, then it will be for future research to develop them.

Acknowledgments. I'd like to thank audience members at the PT-AI 2017 conference, as well as Vincent Müller and the reviewers for this volume. Special gratitude goes to Daniel Kokotajlo and Miriam Johnson for their comments on earlier drafts of this paper.

References

Armstrong, S., Bostrom, N., Shulman, C. (2013). Racing towards the precipice: A model of artificial intelligence development. Technical Report. Future of Humanity Institute. https://www.fhi.ox.ac.uk/wp-content/uploads/Racing-to-the-precipice-a-model-of-artificial-intelligence-development.pdf. Accessed 24 December 2017.

Badhwar, N. (2003). Love. In: LaFollette, H. (ed.) *Practical Ethics*, pp. 42–69. Oxford University Press, Oxford.

Berscheid, E. (2006). Searching for the meaning of "love". In: Sternberg, R. J. & Weis, K. (eds.) *The New Psychology of Love,* pp. 171–183. New Haven, CT: Yale University Press.

Berscheid, E. (2010). Love in the fourth dimension. *Annual Review of Psychology*, 61, 1–25.

Berscheid, E. & Hatfield, E. (1974). A little bit about love. In Huston, T. L. (ed.) *Foundations of Interpersonal Attraction,* pp. 355–381. New York: Academic Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, Oxford.

Button, C. M., & Collier, D. R. (1991). *A comparison of people's concepts of love and romantic love.* Paper presented at the Canadian Psychological Association Conference, Calgary, Alberta.

Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-10), 7–65.

Dyck, A. (1968). Referent-models of loving: A philosophical and theological analysis of love in ethical theory and moral practice. *Harvard Theological Review,* 61(4), 525–545.

Fehr, Beverly. (2013). Social psychology of love. In: Simpson, J. & Campbell, L. (eds.) *The Oxford Handbook of Close Relationships.* Oxford: Oxford University Press.

Fehr, B. & Russell, J. A. (1991). The concept of love viewed from a prototype perspective. *Journal of Personality and Social Psychology*, 60, 425–438.

Fletcher, Guy (2016). *The Routledge Handbook of Philosophy of Well-being*. New York: Routledge.

Frankfurt, H. (1999). Autonomy, necessity, and love. In: *Necessity, Volition, and Love*, pp. 129–41. Cambridge University Press, Cambridge.

Fredrickson, B. (2016). Love: Positivity resonance as a fresh, evidence-based perspective on an age-old topic. In: Barrett, L., Lewis, M., Haviland, J. (eds.) *Handbook of Emotions,* 4th edition, pp. 847-858. Guilford Press, New York.

Gallie, W. B. (1955). Essentially contested concepts. *Proceedings of the Aristotelian Society,* 56, 167-198.

Good I. (1965). Speculations concerning the first ultraintelligent machine. *Advances in Computers,* 6, 31–88.

Harris, C. (1976). Love as the basic moral principle in Paul Ramsey's ethics. *Journal of Religious Ethics,* 4(2), 239–258.

Hegi, K., Bergner, R. (2010). What is love? An empirically-based essentialist account. *Journal of Social and Personal Relationships,* 27(5), 620–636.

Helm, B. (2017). Love. In: Zalta, E. (ed.) *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/love/. Accessed 30 March 2018.

Kline, S. L., Horton, B., & Zhang, S. (2008). Communicating love: Comparisons between American and East Asian university students. International Journal of Intercultural Relations, 32(3), 200–214.

Luby, V., & Aron, A. (1990). *A prototype structuring of love, like, and being-in-love*. Paper presented at the Fifth International Conference on Personal Relationships, Oxford, UK.

Merli, D. (2009). Possessing moral concepts. *Philosophia,* 37, 535–556.

Müller, V. & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In: Müller, V. (ed.) *Fundamental Issues of Artificial Intelligence, pp.* 553-571. Berlin: Springer.

Picard, R. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Ramsey, P. (1950). *Basic Christian Ethics*. Charles Scribners Sons, New York.

Ryan, R. & Deci, E. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78.

Scheutz, M. (2014). Artificial emotions and machine consciousness. *Cambridge Handbook of Artificial Intelligence.* Cambridge: Cambridge University Press.

Singer, I. (2009). *Philosophy of Love: A Partial Summing-up*. MIT Press, Cambridge, MA.

Slote, M. (1983). *Goods and Virtues.* Clarendon Press, Oxford.

Sloman, A. and Croucher, M. (1981). Why robots will have emotions, in *Proceedings of the 7th International Joint Conference on AI* (pp. 197–202).

Snyder, C. R. & Lopez, Shane J. (2009). *The Oxford Handbook of Positive Psychology.* Oxford: Oxford University Press.

Soares, Nate & Fallenstein, Benja (2015). Aligning superintelligence with human interests: A technical research agenda. Machine Intelligence Research Institute. https://intelligence.org/files/TechnicalAgenda.pdf. Accessed March 10, 2018.

Velleman, D. (1999). Love as a moral emotion. *Ethics,* 109(2), 338–374.

Wolf, S. (1992). Morality and partiality. *Philosophical Perspectives,* 6, 243–259.

Yampolskiy, Roman (2016). *Artificial Superintelligence: A Futuristic Approach.* New York: CRC Press.

Yudkowsky, Eliezer (2004). Coherent extrapolated volition. Machine Intelligence Research Institute. https://intelligence.org/files/CEV.pdf. Accessed 30 December 30 2017.