Philosophical Studies
https://doi.org/10.1007/s11098-024-02212-9

Promotionalism, orthogonality, and instrumental convergence

Nathaniel Sharadin[1]

Accepted: 18 August 2024

Abstract

Suppose there are no in-principle restrictions on the contents of arbitrarily intelligent agents' goals. According to "instrumental convergence" arguments, potentially scary things follow. I do two things in this paper. First, focusing on the influential version of the instrumental convergence argument due to Nick Bostrom, I explain why such arguments require an account of "promotion", i.e., an account of what it is to "promote" a goal. Then, I consider whether extant accounts of promotion in the literature—in particular, *probabilistic* and *fit-based* views of promotion—can be used to support dangerous instrumental convergence. I argue that neither account of promotion can do the work. The opposite is true: accepting either account of promotion undermines support for instrumental convergence arguments' existentially worrying conclusions. The conclusion is that we needn't be scared—at least not because of arguments concerning instrumental convergence.

Keywords Instrumental convergence · Orthogonality · Promotion · Instrumental rationality · Artificial intelligence · Existential risk

✉ Nathaniel Sharadin
natesharadin@gmail.com

[1] Department of Philosophy, University of Hong Kong, Hong Kong, China

Published online: 21 October 2024

1 Introduction: promotionalism and convergent instrumental reasons

Sometimes, when we do, think, or feel a certain way we promote something. Call someone a *promotionalist* about reasons to act (think, feel) a certain way if they hold that whether there is a reason to act (think, feel)[1] that way depends on whether so acting promotes a suitable object of promotion. Almost everyone is a promotionalist of this kind about *instrumental reasons*: whether there is an instrumental reason for agents to do something depends on whether their doing it promotes the achievement of one of their goals.[2]

For instance: whether there is an instrumental reason for me to put up my umbrella depends on whether doing so promotes one of my goals, such as the goal of remaining dry. Suppose it does not promote my goal because my umbrella is full of holes. Then, there is no instrumental reason to put up my umbrella. Or suppose it does not promote my goal because I have no such goal: I instead want to get wet. Then, again, there is no instrumental reason to put up my umbrella. Finally, suppose that I aim to stay dry and my umbrella is *not* full of holes. Then, there is intuitively an instrumental reason to put up my umbrella, since doing so promotes my goal of staying dry.[3]

According to a different idea, there are no restrictions on the content of intelligent agents' goals. Nick Bostrom calls this idea the *orthogonality thesis* and describes it like this:

> Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.[4]

Putting ideas together, it's tempting to conclude we can't say anything informative about arbitrarily intelligent agents' instrumental reasons. For, an agent's instrumental reasons are a function of whatever promotes the achievement of their goals

[1] Going forward, I'll mostly be talking about reasons for action. But what I say should go, too, for reasons to think, feel, and comport our minds in other ways, assuming there are reasons for these things. For relevant discussion, see (Sharadin, 2015b).

[2] This is sometimes taken to be definitional of instrumental reasons. See (Broome, 2013) for discussion. Promotionalism about other kinds of reasons is commonplace, too. For example, there are promotionalists about epistemic reasons (Kornblith 1993; Sharadin, 2019; 2018; Cowie, 2014); moral reasons (Schroeder, 2007); prudential reasons (Dorsey, 2012; Dale Dorsey, 2021); and aesthetic reasons (Whiting, 2023). I'll sometimes say that an action "promotes a goal" rather than "promotes the achievement of a goal," but I intend these to be read as meaning the same.

[3] Going forward, I'll ignore differences between agents' goals, desire, ends, and other cognate notions. I assume each can be the suitable object of promotion on which an agent's instrumental reasons depends. I'll focus on "goals" because this is the particular object of promotion picked out by the particular instrumental convergence arguments I am interested in in this paper. This is a harmless terminological stipulation.

[4] (Bostrom, 2012, 73). Here, Bostrom means "final" goals to contrast with (merely) "instrumental" goals in the same way that "final" value is typically contrasted with "instrumental" value. For discussion, see (Korsgaard, 1983).

(promotionalism about instrumental reasons), and those goals could be any way whatsoever (the orthogonality thesis).

This conclusion would be too quick. For, there may be actions that promote the achievement of agents' goals (almost) whatever their specific content. Then, (almost) any agent would have instrumental reason in favor of those actions. To illustrate with an extreme example, suppose there is a magic button, and that pressing the button has one effect: if an agent presses it, then one of their goals is immediately achieved. If such a button exists, then there is an instrumental reason for each agent to press it (and press it, and press it…), no matter what the specific content of those agents' goals.[5]

With some hesitation, I'll follow Bostrom and call reasons that have this feature "convergent" instrumental reasons.[6] Convergent instrumental reasons are instrumental reasons to act (think, want, feel, behave) in particular ways that we can reasonably expect agents to have, (almost) regardless of the specific content of those agents' goals. They are the reasons instrumentally rational agents will (almost) always "converge" on.[7] The idea of convergent instrumental reasons isn't a new one. In a moment I'll explain the use of this idea that's my focus in this paper. Before that, let's very quickly look at five uses of convergent instrumental reasons in a variety of philosophical arguments. This will help frame the discussion to come.

First, consider Kant's argument that there is a reason to (sometimes) help others when they are in need, i.e., an "imperfect" duty to be "beneficent."[8] According to one common interpretation, Kant's argument depends on thinking that, whatever the specific content of their goals, physically and cognitively limited agents will sometimes need others' help to secure achieve those goals. Hence, there is an (instrumental) reason to want to live in a world where people are disposed to (at least sometimes) help one another.[9]

Second, consider John Rawls's idea of "primary goods."[10] According to Rawls, primary goods are those goods that a rational and reasonable person exercising their

[5] There are some technical complications that mean this may not be strictly true, depending on how we choose to formalize things. But we can safely ignore these issues at present. We'll return to some of them below, in § 3.

[6] The hesitation is because, typically, such reasons are referred to as "universal" reasons, or, sometimes even more confusingly, "categorical" reasons. For example, see the discussions in (Schroeder, 2007 esp. Chapter 5; Sharadin, 2018; Kornblith, 1993; Korsgaard, 1996). But: a literature has developed around the idea of convergent instrumental reasons in response to Bostrom's arguments (c.f. Gallow, 2024; Ngo, Chan, & Mindermann 2022; Grace, 2022; Drexler, 2019; Christian, 2020), so it makes sense to play out the argument in those terms.

[7] To be clear: Bostrom's claim concerning instrumentally convergent reasons is about what happens in a wide range of circumstances for a wide range of final goals, but not for literally all situations and for literally all possible goals. Hence it's not sufficient to show that Bostrom's argument is mistaken to (simply) show the relatively straightforward fact that there is some possible goal such that if an agent were to have that goal they wouldn't have instrumentally convergent reason to engage in dangerous behavior. I discuss this issue in more detail below, in § 3.4. Thanks to an anonymous referee for urging clarity on this point.

[8] (Kant, 1785).

[9] For discussion, see (Rawls, 2000, 172–76, 234.).

[10] For Rawls's discussion, see (Rawls, 1971; 1999, 54; 2001).

"capacity for a conception of the good," i.e., the ability to have, revise, and pursue goals, would want, whatever the specific conception of the good they in fact have.[11] Rawls uses the notion of primary goods in an argument for his two principles of justice; in particular, he thinks that an interest in the primary goods means that parties negotiating on principles of justice behind the veil of ignorance would select minimax principles for the distribution of economic surplus and a rights-based system guaranteeing certain inalienable protections for citizens.[12] In short: all (relevant) agents have instrumental reason to secure (a minimally acceptable distribution of) primary goods in deciding on political arrangements, whatever the specific content of their actual goals (what Rawls calls their "conception of the good").

Third, consider Christine Korsgaard's idea that, whatever their specific goals, all rational agents have an instrumental reason to value their own capacity for *having goals*.[13] Korsgaard uses this result to argue that a commitment to morality is a constitutive component of agency as such.[14] That argument has been influential but controversial. Whatever the merits of Korsgaard's argument, the idea is the same: all agents have reason to do something (value their own capacity for having goals), whatever the specific content of their actual goals.

Fourth, consider Mark Schroeder's neo-Humean account of reasons. Schroeder explicitly accepts promotionalism about instrumental reasons and instrumentalism about practical reasons, i.e., the view that practical reasons are all instrumental reasons (he calls this view "hypotheticalism"). [15] As we've seen, this seems to imply that there are no facts about what all agents have reason to do, given that they might want to achieve anything at all. But this, in turn, seems to preclude the possibility of *moral* reasons, since moral reasons are typically taken to be reasons all agents have to do various things.[16] Since Schroeder wants to allow for the possibility of moral reasons in his account, he points out that there could in principle exist things the doing of which would promote agents' goals whatever those goals' specific content.[17]

Schroeder's use of the idea of convergent instrumental reasons is instructive. We can imagine other kinds of views that accept the combination of promotionalism about instrumental reasons and instrumentalism about some further domain of reasons. Those views will face symmetrical challenges insofar as they're interested in accounting for convergence in the further kind of reason.

So, fifth and finally, consider epistemic instrumentalism. Epistemic instrumentalists are promotionalists about instrumental reasons, and they are instrumentalists about epistemic reasons.[18] Hence, they face a version of the problem Schroeder faces with moral reasons with respect to epistemic reasons. Whether there is a reason to

[11] (Rawls, 2001, 58 and following).
[12] (Rawls, 2001 esp the argument stressing the "first condition" in III).
[13] (Korsgaard, 1996).
[14] See especially the argument in (Korsgaard, 2009).
[15] (Schroeder, 2007).
[16] For discussion, see (Schroeder, 2007 esp, chapter 6).
[17] (Schroeder, 2007 esp, 6.2).
[18] (Sharadin, 2022).

believe in accordance with the evidence intuitively shouldn't depend on the specific content of an agent's goals. And epistemic instrumentalists reply that, happily, it doesn't: this is because believing in accordance with the evidence is a useful way of achieving one's goals (almost) no matter what, precisely, those goals comprise.[19] Hence, there is (almost) always an epistemic (albeit instrumental) reason to believe in accord with the evidence.[20]

As I indicated, these five uses of the same idea are each controversial in their own ways. But they have two important things in common. First, they all appeal to the thought that we can say something substantive about agents' instrumental reasons without saying anything (quite so) substantive about the content of their goals.

Second, all these argumentative uses of convergent instrumental reasons are *optimistic* in the following sense: they are arguments in favor of agents having reasons to do things that we can all independently agree are, on the whole, *good*. They are reasons to be beneficent (Kant), to cooperate (Rawls), to respect humanity (Korsgaard), to do what's moral (Schroeder), and to believe what's true, or what the evidence supports (epistemic instrumentalism). They are in that way optimistic takes on agents' convergent instrumental reasons.

Nick Bostrom is not optimistic.

According to Bostrom, the existence of specific convergent instrumental reasons increases the likelihood of the literal apocalypse. How is this supposed to be possible? Here is the basic idea: first, imagine an agent that is "super" intelligent, and also potentially very capable or powerful in ways that are not immediately transparent to us.[21] Next, suppose this agent has goals with potentially very alien content (as per the orthogonality thesis). Now ask: What would a powerful, potentially very intelligent agent like this have reason to do, (almost) whatever they want to achieve? The answer, worryingly, might be: acquire more power.[22] Or: acquire the means of acquiring power. Or: acquire all the resources inside their lightcone.[23] From here, Bostrom and others who accept these arguments think it's only a few theoretical steps to the literal apocalypse. And it's not hard to see why: superintelligent instrumentally rational agents with instrumental reasons in favor of single-mindedly

[19] See (Kornblith, 1993; Cowie 2014; Côté-Bouchard, 2015; Sharadin, 2018; Sharadin, 2022; Rinard, 2015; 2017; 2019).

[20] More recently, social epistemic instrumentalists have argued that the relevant aims being promoted are the aims of groups, or collectives. See (Hannon and Woodard, n.d.; Dyke, n.d.).

[21] Here, I do not attempt to specify exactly what superintelligence comprises, nor do I engage with critiques of this framing. For an overview, see (Mitchell, 2019). Instead, I am arguing against a particular view that accepts, very broadly, the account of "superintelligence" given in (Bostrom, 2014). Nothing in my argument relies on any particular conception of superintelligence, since I target a claim about instrumentally rational agents' instrumental reasons that doesn't depend on agents being "super" rational or "super" instrumentally rational, or any other feature of intelligence reaching beyond human-level capabilities. Going forward, I therefore mostly ignore the idea of "superintelligence" and just talk about more or less "capable" agents.

[22] For concerns about power-seeking AI, see (Carlsmith, 2022 ; Turner et al. 2021).

[23] For variations on these kinds of argument, that instrumental convergence (may) lead to existential risk, see (Carlsmith, 2022; Omohundro et al., 2007; The Basic AI Drives, 2007; Shulman, 2010; Turner et al. 2021; Hendrycks, 2023; Hendrycks, Mazeika, and Woodside 2023). For general discussion of arguments for existential risk from advanced AI, see (Bales, D'Alessandro, and Kirk-Giannini 2024).

pursuing the maximization of their own power is arguably not at all safe for humanity.

Setting aside the details of the argument for the moment, the conclusion of this line of reasoning should strike us as deeply worrying, and it will strike many of us as surprising. We might have thought, as Kant, Rawls, et. al. seem to think, that the arc of instrumental reasons bends toward justice – or at least toward relatively good outcomes.[24] But if Bostrom is right, then this is based on an anthropocentric bias – we're simply assuming that artificial systems will have goals suitably similar to human ones. But once we correct for that bias by noticing that an artificial system could well have any set of goals whatsoever (per the orthogonality thesis) we get a much more worrying result. Whereas it might be true that *human beings* will (typically) have instrumental reasons to cooperate with one another, very capable artificial systems may well have instrumental reasons to do things that are very dangerous for humanity, such as acquire all the resources within their lightcone. This is because, perhaps surprisingly, their doing these dangerous things would promote their goals, almost whatever those goals' content. If true, this is a worrying, pessimistic conclusion.

The primary aim of this paper is to argue for two claims. The first claim is that the success of arguments of this sort—so-called "instrumental convergence arguments" depends on what account of *promotion* we accept. The second claim is that extant accounts of promotion from the philosophical literature—in particular, *probabilistic* and *fit-based* accounts of promotion—do not in fact support these arguments. The conclusion of the paper is therefore that, absent some alternative account of promotion that delivers the troubling results, we should not be worried by arguments concerning instrumental convergence.

Here is the plan for the remainder of the paper. §2 gives a more careful overview of Bostrom's argument, focusing on the role of an account of promotion in delivering his pessimistic conclusion. §3 considers whether a *probabilistic* account of promotion can be used in this argument and argues that it cannot. §4 considers whether a *fit-based* account of promotion can be used in this argument and argues that it also cannot. §4 concludes.

2 Bostrom's argument & dangerous convergent promotion

In this section I'll lay out Bostrom's argument concerning dangerous instrumental convergence in more detail and explain why its pessimistic conclusion relies on an account of what it means to promote a goal. To that end, it's worth beginning by quoting Bostrom in full and at length:

> The instrumental convergence thesis suggests that we cannot blithely assume that a superintelligence with the final goal of calculating the decimals of pi

[24] Or at least: good outcomes when embedded within the right political institutions. For discussion, see (Rawls, 2001).

> (or making paperclips, or counting grains of sand) would limit its activities in such a way as to not materially infringe on human interests. An agent with such a final goal would have a convergent instrumental reason, in many situations, to act to acquire an unlimited amount of physical resources and, if possible, to eliminate potential threats to itself and its goal system. It might be possible to set up a situation in which the optimal way for the agent to pursue these instrumental values (and thereby its final goals) is by promoting human welfare, acting morally, or serving some beneficial purpose as intended by its creators. However, if and when such an agent finds itself in a different situation, one in which it expects a greater number of decimals of pi to be calculated if it destroys the human species than if it continues to act cooperatively, its behavior would instantly take a sinister turn. This indicates a danger in relying on instrumental values as a guarantor of safe conduct in future artificial agents that are intended to become superintelligent and that might be able to leverage their superintelligence into extreme levels power and influence. (Bostrom, 2012, 84).

Let's unpack this. Remember, convergent instrumental reasons are reasons agents have regardless of their goals (e.g. to count grains of sand, or calculate decimals of pi, etc.); here, Bostrom's claim is that superintelligent agents may have convergent instrumental reasons to do dangerous actions such as those involved in acting to acquire "extreme levels" of physical resources.[25] Call this claim, which is the worrying, pessimistic conclusion of the argument:

> **Dangerous Convergent Instrumental Reason**: Almost whatever a superintelligent agent's goals, there is an instrumental reason for that agent to act to acquire extreme levels of physical resources.

Dangerous Convergent Instrumental Reason is Bostrom's pessimistic conclusion analogous to Rawls, et. al's optimistic conclusions about humans' convergent instrumental reasons to (e.g.) coordinate on principles of justice. What is Bostrom's argument for Dangerous Convergent Instrumental Reason? Here is Bostrom's idea. Physical resources are intuitively useful for achieving a wide variety of different goals. So, intuitively at least, acting to acquire "extreme levels" of these resources can be a way of *promoting* the achievement of a wide variety of goals for those agents.[26] Call this claim:

> **Dangerous Convergent Promotion**: Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources promotes those goals.

[25] Here and throughout the next two sections, I focus on the specific action of acting to acquire extreme levels of physical resources. The argument I present can be adjusted, *mutatis mutandis*, for other cases of potentially dangerous behavior, such as acting to acquire extreme levels of power, or extreme levels of influence.

[26] Thanks to [removed for blind review] for suggesting clarity on this point.

This is not by itself enough to complete the argument. In order to get from Dangerous Convergent Promotion to Dangerous Convergent Instrumental Reason, we need a linking principle between facts about *promotion* and facts about agents' *instrumental reasons*. Promotionalism about instrumental reasons is more than sufficient.[27]

Promotionalism about Instrumental Reasons: There is an instrumental reason for an agent to φ iff (and because) φ-ing promotes one of their goals.[28]

Together, Dangerous Convergent Promotion and Promotionalism about Instrumental Reasons deliver the worrying result that is Dangerous Convergent Instrumental Reason. Here, then, is Bostrom's argument in premise-conclusion form:

(1) There is an instrumental reason for a superintelligent agent to act to acquire extreme levels of physical resources iff (and because) doing so promotes one of their goals. (**Promotionalism about Instrumental Reasons [PIR]**)
(2) Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources promotes those goals. (**Dangerous Convergent Promotion [DCP]**)
(3) So: Almost whatever a superintelligent agent's goals, there is an instrumental reason for that agent to act to acquire extreme levels of physical resources. (**Dangerous Convergent Instrumental Reason [DCIR]**).

This argument is valid. (1) Follows from a very widely accepted view about the nature of instrumental reasons, viz. PIR. In what follows, I'll assume PIR is true, and so that (1) is also true. (3) follows from (1) and (2). Moreover, it doesn't strike me as an exaggeration to call the reason identified in (3) a *dangerous* convergent instrumental reason.[29] Again, there are many different possible scenarios. But intuitively, sometimes at least, acting to acquire extreme levels of physical resources could involve excluding human beings from access to those resources. That could clearly be dangerous for particular humans, or even for humanity itself. Here, I assume this is correct: if there is a reason of the sort identified in (3) then it is suitably "dangerous" for humans (or humanity).

Hence the only remaining question is whether we should accept (2): Does acting to acquire extreme levels of physical resources in fact promote a superintelligent agent's goals, almost whatever those goals' specific content?

[27] It's clear from what Bostrom writes that he accepts PIR. See especially (Bostrom, 2014, esp. chapters 6, 7, 12).

[28] It's clear from what Bostrom writes that he accepts PIR. See especially (Bostrom, 2014, esp. chapters 6, 7, 12).

[29] One idea is that this is not dangerous because such reasons will usually be extremely weak. I think this is likely true; here, I follow the assumption in the literature and ask just whether there *exist* such reasons and ignore the question of their (relative) weight. In a companion paper (N. Sharadin 2024) I argue that even if dangerous convergent instrumental reasons exist, they are likely very weak — potentially *maximally* weak.

It can seem intuitively obvious that acting to acquire extreme levels of physical resources does in fact promote an agent's goals, almost whatever those goals' specific content. To illustrate, take one of the most useful physical resources: *energy*. Energy (e.g. in the form of electricity) is a very useful thing to have (almost) no matter what one's goals, since it can be directly used in the achievement of one's goals, and excess energy can be sold or traded, in order to produce (acquire) more all-purpose means (e.g. currency) for achieving one's goals, again no matter what, specifically, those might be. Moreover, at least intuitively, there doesn't appear to be any limit to the usefulness of acting to acquire more (and more, and more) physical resources of this kind, no matter what one's goals, at least assuming that excess quantities of those resources will be exchangeable for some other useful goods, or can be traded for a store of value (such as exchangeable currency).

Hence, (almost) no matter what the specific content of one's goals, it can seem obviously true that acting to acquire an additional unit of physical resources promotes the attainment of that goal. After all, even if one doesn't directly need more (e.g.) joules, one can always trade excess joules for (e.g.) jewels, and then in turn trade jewels for whatever turns out to be useful for promoting one's goals. In effect, the natural, intuitive idea is that more all-purpose means are always good, from the point of view of promoting one's goals.

As I said, I think this idea is intuitive. But it is very important not to leave things at an intuitive level. The conclusion of the argument, that superintelligent agents may have instrumental reason to do things that are dangerous for humans or for humanity itself is not something we should accept on the basis of intuitive judgments about cases. For, if it's true that sufficiently capable artificial agents will (almost all) have instrumental reason to act to acquire physical resources without limit, then maybe we shouldn't build such things in the first place. Or maybe only governments should be entrusted with the development of (sufficiently capable) artificial systems. Or maybe it's time to build a bunker. These are not courses of action or policies we should undertake on the basis of intuitive judgments about what promotes what.

Instead, we are due a principled account of what it means to say that an action *promotes* a goal. Only with such an account in-hand can we systematically evaluate the claim that there is (or might be) Dangerous Convergent Promotion. In the remainder of this paper, I argue that extant accounts of promotion do not in fact support the claim that there is Dangerous Convergent Promotion. Therefore, without some alternative account of promotion that does in fact support Dangerous Convergent Promotion, we should not accept Bostrom's argument or others that, like it, rely on Dangerous Convergent Promotion.

3 Dangerous convergent probabilistic promotion?

As we just saw, the crucial premise in Bostrom's argument is (2):

Dangerous Convergent Promotion (DCP): Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources promotes those goals.

Above, we saw that DCP is intuitive – that was what the example of the endless acquisition of (e.g.) energy illustrated. But setting aside intuitive judgments about cases, whether DCP is true depends on what, precisely, it takes for an action to "promote" an agent's goals.

In the literature on the nature of promotion, there are two broad families of views: *probabilistic* accounts, and *fit-based* accounts. In the next two sections (§§ 3–3), I'll argue that neither a probabilistic account of promotion nor a fit-based account of promotion delivers Dangerous Convergent Promotion. I'll focus on probabilistic accounts in this section. Then (§4) I turn to fit-based accounts.

3.1 Probabilism about promotion

According to probabilists about promotion, an action's promoting a goal is a matter of the action's increasing the *probability* of that goal's achievement (relative to some baseline). Probabilists about promotion all accept some instance of the schema:

Probabilism about Promotion: Agent A's φ-ing promotes goal G iff pr(G|A φs) > B.

Where B (for *baseline*) is then filled in by the specific probabilistic account of promotion on offer; variants on probabilistic accounts of promotion abound.[30] The idea behind probabilism is intuitive. To illustrate the basic mechanics of the view, suppose the relevant baseline is: the probability of the goal being achieved given that the agent does nothing at all (i.e., B = pr(G | A does nothing).[31] Now suppose as before that I want to stay dry. Then, according to probabilism about promotion, opening my umbrella promotes my goal of staying dry just in case the probability of my staying dry given that I open the umbrella is greater than the probability that I stay dry given that I do nothing. This probabilistic account of promotion therefore delivers intuitive results, at least in a wide range of cases. If I'm under a canopy, then opening my umbrella doesn't promote my goal of staying dry (doing nothing doesn't reduce my chances of staying dry). If I'm outside and it's raining, then it does (doing nothing makes it more likely I'll get wet). If the umbrella has holes, it doesn't. And so on.

Some of Bostrom's remarks suggest that he is a probabilist of one kind or another about promotion. Here, my aim here isn't to engage in Bostrom exegesis, but it's worth sampling a few relevant passages:

[30] See, for example, (Coates, 2013; Elson, 2019; Finlay, 2006; 2014; Lin, 2018; Schroeder, 2007).

[31] This is the baseline suggested by (Schroeder, 2007). I return to the problem of selecting a suitable baseline below, in § 3.2.

> "Several instrumental values can be identified which are **convergent in the sense that their attainment would increase the chances of the agent's goal being realized** for a wide range of final goals and a wide range of situations" [...] "in many scenarios there will be future **actions [an agent] could perform to increase the probability of achieving its goals**" [...] "There are special situations in which **cognitive enhancement may result in an enormous increase in an agent's ability to achieve its final goals**" (Bostrom, 2014, 132, 132, 134, emphasis added).

Each of these claims suggests that Bostrom is thinking about promotion in terms of the schema above, i.e., in terms of an agent's action making a positive difference to the *probability* of achieving some goal relative to a baseline. As we've already seen, it's important not to rely on our intuitive judgments about promotion when it comes to this argument. So, in order to evaluate these claims we require some systematic account of the baseline relative to which the action of (e.g.) acting to acquire extreme levels of physical resources is supposed to increase the probability of achieving an agent's goals in order to count as *promoting* those goals.

The literature on probabilism about promotion illustrates that it has proven extremely, perhaps surprisingly difficult to identify a suitable baseline relative to which the likelihood of an agent's achieving some goal given that they perform an action must go up, in order for the agent to count as doing something that *promotes* that goal. There are roughly three views of the baseline extant in the literature. The difficulties faced by these three views, and the remedy for them recently proposed in the literature, is instructive in the present context. It's therefore worth quickly working through the problems with these accounts. I'll then discuss the lesson we learn by reflecting on these problems, and then explain the upshot for Bostrom's argument.

3.2 Baseline problems for probabilism

Consider first the baseline we discussed above, viz. the agent's *doing nothing*, so that we have:

Agent A's φ-ing promotes goal G iff pr(G|A φs) > pr(G | A does nothing).[32]

The problem with this account of the baseline is that it systematically undercounts cases of promotion.[33] Suppose Artie wants a job at Princeton and applies for it. Now suppose that if Artie does anything at all (e.g., follows up about the position), they will be rejected from the applicant pool (how gauche). If instead they do nothing at all, they will get the job (the fix is in). Then, doing nothing is what promotes Artie getting the job. But the baseline just articulated doesn't allow for this obvious fact. After all, the probability that Artie gets the job given that Artie does nothing is the

[32] (Schroeder, 2007).

[33] (Evers, 2009) appears to be first to have noticed this kind of problem with probabilistic promotion views. It was further developed by (Behrends and DiPaolo, 2011). For extended critical discussion, see (Sharadin, 2015a; Behrends and DiPaolo 2016; N. Sharadin 2016; Coates, 2013; Lin, 2018; Elson, 2019).

same as the probability that Artie gets the job given that Artie does nothing. Hence doing nothing can't promote Artie getting the job. But doing nothing obviously *can* promote Artie getting the job: in this case, it's the *only way* for Artie to promote it!

Consider next the baseline suggested by Stephen Finlay: *not performing the action in question*, which gets us [34]

Agent A's φ-ing promotes goal G iff pr(G|A φs) > pr(G | A doesn't φ).

This is also incorrect, and for related reasons. Suppose the initial setup is as before. But suppose this time the norms involving following up on a job application are somewhat less insane, and that there are three things Artie could do. Artie could write an email to the chair of the department, write an email to the dean of the school, or write an email to the president of the college. If Artie writes an email to either the chair or the dean, they'll get the job (chairs and deans are pushovers). If Artie writes an email to the president, they won't get the job (the president's a jerk). Finally, suppose if Artie doesn't write an email to the chair, they'll write to the dean. Then, the baseline just articulated entails that Artie cannot promote their end of getting the job by writing to the chair. This is because if Artie doesn't write to the chair they'll write to the dean. But the probability of getting the job given that they don't write to the chair (and instead write to the dean) is exactly the same as the probability of getting the job given that they write to the chair – in either case, they are sure to get the job. This is clearly wrong. Given the details of the case, Artie can promote their goal of getting the job by writing to the chair: in fact, writing to the chair *guarantees* they'll get the job! Every plausible account of promotion should say that *guaranteeing* a goal is achieved is a way of promoting that goal.

Finally, consider a *temporal* baseline, according to which:

Agent A's φ-ing promotes goal G iff for time t_0 immediately prior to an agent A's φ-ing at t_1 and later time t_2 immediately after A's φ-ing, $pr_{t2}(G) > pr_{t0}(G)$.[35]

Intuitively, φ-ing promotes a goal when it causes the probability of the goal's achievement to go up. But this is also incorrect. Suppose Artie is deciding between regimes for publishing their papers in an effort to secure tenure (they got the job!), and they're considering adopting one of three plans at t_1[36]

Regular Publication (R): Regularly publish papers

Sporadic Publication (S): Sporadically publish papers

No Publication (N): Never publish papers

Artie is conscientious. So, there's a good chance they'll adopt a regular publication schedule; but there's some chance they'll decide to publish sporadically, and a very small chance they will opt not to publish at all. In particular, suppose that the following probabilities hold at t_1 with respect to Artie's likelihood of adopting each of these regimes:

$$pr_{t1}(R) = .88$$
$$pr_{t1}(S) = .1$$

[34] (c.f. Schroeder, 2007).

[35] (c.f. Lin, 2018).

[36] (c.f. Lin, 2018).

$pr_{t1}(N) = .02$

Finally, suppose the following conditional probabilities hold with respect to the chances of Artie's securing tenure (G) given the various publication regimes:

$pr_{t1}(G|R) = .95$
$pr_{t1}(G|S) = .75$
$pr_{t1}(G|N) = .001$

These probabilities reflect the (perhaps lamentable) fact that people who adopt regular publication schedules are somewhat more likely to secure tenure than those who adopt sporadic ones (though those who sporadically publish are still quite likely to secure tenure), and both groups are much more likely to secure tenure than those who opt not to publish at all. In order to decide what courses of action could promote Artie's end of getting the job we need to know the antecedent probability of their getting tenure at t_1 so that we can compare it to the chances of their tenure given that they adopt one of the publication regimes. Assuming the regimes are a partition of possible behaviors for Artie, we can calculate that probability:

$pr_{t1}(G) = pr_{t1}(R) * pr_{t1}(G|R) + pr_{t1}(S) * pr_{t1}(G|S) + pr_{t1}(N) * pr_{t1}(G|N) = .88 * .95 + .1 * .75 + .02 * .001 \approx .911$

But then, according to the view on offer, the only available action that Artie could perform at t_1 that would promote their goal of getting tenure is to begin a regular publication regime (R). If instead Artie begins publishing sporadically (S), then the probability of getting tenure is equal to the prior probability of getting tenure given that they publish sporadically (i.e., 0.75). But then this view entails that sporadically publishing not only doesn't *promote* Artie's goal of getting tenure, it *dispromotes* it.[37] But that is absurd: even if sporadically publishing doesn't *best* promote Artie's goal of getting tenure, it doesn't *dispromote* it! In the language of reasons: Artie has *most* or *strongest* instrumental reason to publish regularly, but at least *some* or *some weighty* instrumental reason to publish sporadically. (Artie has *least* or *weakest* reason to publish not at all.)

Selecting a suitable baseline for probabilistic promotion is therefore an extremely tricky matter. In the next subsection (§3.3), I explain the recent suggestion offered in the literature designed to solve these problems. Then (§3.4), I explain why adopting this solution vitiates the case for Dangerous Convergent Promotion. The result is that a probabilistic account of promotion can't do the work required by Bostrom's argument.

[37] The language of 'dispromotion' is clunky. It's tempting to say that the activity would 'frustrate' the goal. But an anonymous referee points out that it's tempting to read 'frustrate' as a success term, such that to frustrate an agent's goal is to (completely) prevent its achievement.

3.3 Contrastive probabilistic promotion

There's a simple, elegant, unified solution to all these (and other) problems with selecting a baseline for a probabilistic account of promotion. Unfortunately, as I'll shortly explain (§3.4), adopting this solution interrupts Bostrom's argument for Dangerous Convergent Promotion. First, let's get the solution on the table.[38]

The solution is to treat promotion as a *contrastive* matter, and so to treat the *baseline* for the probabilistic increase required for promotion as set by a relevant contrast class. So, rather than taking promotion to be a *non*-contrastive matter, whereby an action either promotes (or fails to promote) a goal *simpliciter*, we should take promotion to be a contrastive notion, according to which an action either promotes, or fails to promote, a goal *as compared to other actions*. Working within the basic probabilistic framework, we replace:

> Probabilism about Promotion: Agent A's φ-ing promotes goal G iff pr(G|A φs) > B.

with:

> Contrastive Probabilism about Promotion: Agent A's φ-ing rather than ψ-ing promotes goal G iff pr(G|A φs) > pr(G|A ψs).

Note the two differences between Contrastive Probabilism and its non-contrastive counterpart. First, an agent's action promotes a goal only as contrasted with another action. Second, the baseline against which an agent's action must raise the probability of a goal's achievement in order to count as promoting it is replaced with the probability of the goal's achievement given the contrasting action.

This simple, elegant solution resolves all the baseline problems that we've noted (§3.2). Here is how. Recall the initial problem illustrated by Artie: the baseline "doing nothing" was unable to account for the fact that Artie's doing nothing *promoted* his goal of getting the job. But there is no barrier to saying this on the present account. For, Artie's doing nothing *rather than anything at all* promotes his goal of getting the job. Equally for the problems with the baseline "not performing the action in question". Remember, there were three options and outcomes for Artie: write an email to the chair (get the job), write an email to the dean (get the job), or write an email to the president (not get the job). The problem was that if Artie didn't write the chair, they'd write the dean; and so, it was impossible to say that his writing the chair promoted getting the job. But given *Contrastive* Probabilism, we can say that:

[38] The proposed solution is due to (Sharadin & Dellsén 2019). For a similar proposal regarding *reasons* (in a related context), see (Snedegar, 2017). Other solutions might be available, too. For instance, (Elson, 2019) defends a non-contrastive "minimal probabilism" about promotion that also solves the problems with (simple) probabilism. Interestingly, as an anonymous referee points out, Elson's view also appears to undermine Bostrom's case for his scary conclusion, since it typically won't turn out that the pursuit of extreme levels of physical resources is the action that outranks all other actions in a (superintelligent) agent's ability set. For reasons of space I leave this discussion to another time.

(1) Artie's writing the chair *rather than the president* promotes getting the job.
(2) Artie's writing the dean *rather than the president* promotes getting the job.
(3) Artie's writing the chair *rather than the dean* doesn't promote getting the job.

Given (1), there's a sense in which Artie's writing the chair promotes their goal of getting the job, and by the same token given (2) there's a sense in which writing the dean promotes their goal of getting the job. But it's intuitive that writing the chair is just as good as writing the dean, and this is captured by (3), which says that writing the chair *rather than the dean* doesn't promote getting the job.

What about the third, *temporal* baseline, and the problematic case involving Artie's attempt to get tenure? Here, too, going contrastive solves things. The problem, as we saw, was that the temporal baseline was unable to account for the fact that embarking on a sporadic publication schedule promoted Artie's goal of getting tenure. Intuitively, this was because Artie was antecedently already very likely to do something that would make it likely that they'd achieve their goal. But given a contrastive understanding of promotion, we can easily identify the way in which Artie's sporadic publication in fact promotes his goal, even under those conditions. Consider:

(4) Artie's sporadic publication *rather than no publication* promotes getting tenure.
(5) Artie's regular publication *rather than no publication* promotes getting tenure.
(6) Artie's sporadic publication *rather than regular publication* does not promote getting tenure.

Given (4), there's a sense in which Artie's sporadic publication promotes getting tenure, in that it it does so *as contrasted with no publication*. Equally, given (5), Artie's regular publication promotes their getting tenure. But also, given (6), there's a sense in which sporadic publication *doesn't* promote getting tenure *as contrasted with regular publication*. Again, these are exactly the correct results.

Note that we sometimes *say* that an action promotes a goal, and leave it at that. But on the present account, these claims about promotion *simpliciter* are always shorthand for contrastive promotion claims, where the contrasting action (or actions) is determined by conversational context by way of (e.g.) familiar maxims.[39] For instance, obeying Grice's maxim of quantity, I might in certain contexts tell Artie that emailing the dean promotes getting the job (full stop), since that is all the information that's needed (and no more).[40] But in other contexts, such as where Artie has asked about the possibility of also emailing the chair, it could be infelicitous to assert the non-contrastive promotion claim. As ever, these issues are delicate. The point here is that there is no barrier to interpreting claims about what promotes what

[39] Compare (Sharadin and Dellsén, 2019). For similar remarks about contrastive notions more generally, see (Sinnott-Armstrong, 2008).
[40] (Grice, 1975).

(*full stop*) on the contrastivist picture: as with other contrastive notions, contrast classes can be determined (and made salient) by conversational context.[41]

The upshot of this is that if we're going to be *probabilists* about promotion, we should be *contrastive* probabilists.[42] The problem with "going contrastive" in the present context, as I'll now explain, is that it undermines the case for Dangerous Convergent Promotion.

3.4 Dangerous convergent contrastive probabilistic promotion?

First, a quick recap. Bostrom's argument relies on the claim we've been calling:

> **Dangerous Convergent Promotion (DCP):** Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources promotes those goals.

We are presently investigating accounts of what it means to "promote" a goal in order to see whether these accounts in fact deliver DCP. I just laid out *probabilism* about promotion (§ 3.1), which says that promoting a goal is a matter of making its achievement more likely. We then saw that difficulties identifying a suitable baseline for probabilism (§ 3.2) motivate the simple, elegant solution of *going contrastive* (§ 3.3) with a probabilistic account of promotion.

Having gone contrastive, what should we now say about Dangerous Convergent Promotion? The first thing we should say is that it is strictly incomplete. This is because Dangerous Convergent Promotion is a statement about promotion *simpliciter* but, as we just saw, claims about whether an action promotes some goal (or indeed a wide range of goals) are best understood, on the probabilistic framework, as claims about whether an action probabilistically promotes a goal *as compared to some contrasting action*. Dangerous Convergent Promotion must therefore be understood as shorthand, for a *contrastive* promotion claim, along the lines of:

> **Dangerous Contrastive Convergent Promotion (DCCP):** Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources *rather than ψ-ing* promotes those goals.

Where ψ is a course of action a superintelligent agent could undertake rather than acting to acquire extreme levels of physical resources. The problem, as I'll now explain, is that values of ψ where the relevant instance of DCCP is true don't seem dangerous, and the ones where the relevant instance seems dangerous don't seem true.

It will help to begin by considering what happens to the general form of Bostrom's argument when we replace Dangerous Convergent Promotion with Dangerous *Contrastive* Convergent Promotion; to that end, consider an analogous argument. Suppose I tell you that (almost) whatever it turns out Brett wants, there's an

[41] For discussion in the context of other contrastive notions, see (Sinnott-Armstrong, 2008; Snedegar, 2014; 2017; Weatherson, 2006).

[42] (N. Sharadin and Dellsén 2019).

instrumental reason for him to cook and eat people. That sounds very worrying: stay away from Brett! But (you might ask), what's my argument for this worrying claim? I respond:

(1B) There is an instrumental reason for Brett to cook and eat people iff (and because) doing so promotes one of Brett's goals.

(2B) Almost whatever Brett's goals, cooking and eating people *rather than cooking and eating poison* promotes those goals.

(3B) So: Almost whatever Brett's goals, there is an instrumental reason for Brett to cook and eat people.

I then justify my argument as follows: (1B) follows from promotionalism about instrumental reasons. (2B) is plausibly true: almost whatever Brett wants, he's more likely to get it if he cooks and eats people rather than cooking and eating poison (unless, of course, he wants to die). (3B) follows from (1B) and (2B).

This argument should leave you nonplussed. It doesn't give you any reason to fear Brett, to think Brett's a bad person, or to think Brett is likely to cook and eat people. At best (worst?), it gives you reason to avoid an instrumentally rational Brett in very specific kinds of potentially very unlikely circumstances – ones where he is (likely to be) facing a forced choice between cooking and eating people and cooking and eating poison. What's going on here?

The lesson is that "going contrastive" about promotion in the way required to make good on a probabilistic account of promotion entails "going contrastive" about instrumental reasons, too.[43] And as we can now clearly see, going contrastive about instrumental reasons changes the upshot of the relevant argument. It's therefore misleading to elide the contrastive component in both the first premise and in the conclusion of the argument. Instead, the argument concerning Brett should be:

(1B*) There is an instrumental reason for Brett to cook and eat people *rather than cooking and eating poison* iff (and because) cooking and eating people rather than cooking and eating poison promotes one of Brett's goals.
(2B) Almost whatever Brett's goals, cooking and eating people *rather than cooking and eating poison* promotes those goals.
(3B*) So: Almost whatever Brett's goals, there is an instrumental reason for Brett to cook and eat people *rather than cooking and eating poison*.

The lesson, to repeat, is that if we're promotionalists about instrumental reasons, and we're contrastivists about promotion, then we're (thereby) contrastivists about instrumental reasons.[44] Making this explicit makes it explicit why the initial argument was perhaps unsettlingly *phrased* but not in fact unsettling. It is simply not unsettling for Brett to have an instrumental reason to cook and eat people *rather*

[43] (Sharadin & Dellsén 2019; Snedegar, 2014) make a similar point.

[44] (Snedegar , 2017;2019; 2014) argues that *all* kinds of reasons (e.g. moral reasons, epistemic reasons, etc.) should be understood *contrastively*. Here, I am suggesting that, insofar as *instrumental* reasons are understood in terms of *promotion*, and *promotion* must be understood contrastively (at least within a probabilistic framework), those reasons are therefore correctly understood contrastively. For related discussion, see (N. Sharadin and Dellsén 2019).

than cooking and eating poison even if he has that reason *no matter what his goals might be*. It sure *sounds* unsettling. But it shouldn't worry us. It should only worry us under very specific conditions.

To repeat: This argument only gives us reason to be existentially worried about Brett if we think Brett is likely to be in situations where he'll be faced with the choice between cooking and eating people and cooking and eating poison. In such cases, (almost) no matter what Brett's goals, it turns out that he'll have instrumental reason to cook and eat people *rather than cook and eat poison*. And if we think that Brett is instrumentally rational, that would be existentially worrying to us people who, presumably, do not want *other* people to have instrumental reason to cook and eat us.

I hope it's clear how this makes a difference to Bostrom's argument. Let's restate Bostrom's argument in full, including all the contrastive components in their appropriate places:

> (1*) There is an instrumental reason for a superintelligent agent to act to acquire extreme levels of physical resources rather than ψ-ing iff (and because) doing so promotes one of their goals.
> (2*) Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources rather than ψ-ing promotes those goals.
> (3*) So: Almost whatever a superintelligent agent's goals, there is an instrumental reason for that agent to act to acquire extreme levels of physical resources rather than ψ-ing.

(1*) follows from promotionalism about instrumental reasons *and* the contrastivist insights about promotion. (2*) is Dangerous Contrastive Convergent Promotion (DCCP). And (3*) is the required update to Dangerous Convergent Instrumental Reason, now understood contrastively. What should we make of this updated version of the argument?

I think we should be much less impressed by it. Certainly, we should be much less worried about its conclusion. This is for the same reason as in Brett's case. More specifically in this case, consider the following: whereas it's possible to think of particular ψ-ings such that (2*) is true, the corresponding instances of (3*) are not dangerous. Moreover, when (3*) is in fact plausibly dangerous, the corresponding instance of (2*) is very unlikely to be true. But in order for the updated version of Bostrom's argument to be existentially worrying, both things must be true simultaneously: we must be able to find a case where there is some particular ψ-ing such that the corresponding instance of (2*) is *true*, and the corresponding instance of (3*) is existentially dreadful. In other words, it has to be true *both* that the agent in fact has a reason to acquire extreme levels of resources *rather than* ψ for some value of ψ, and that their having that contrastive instrumental reason is itself existentially risky. But, I claim, we don't have any reason to expect this is true. Let's work through an example to help illustrate the point.

Consider first:

> (2-None) Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources *rather than acting to acquire no resources at all* promotes those goals.

(2-None) is very plausible. Almost no matter what an agent – superintelligent or not! – aims to achieve, acting to acquire extreme levels of resources *rather than acting to acquire no resources at all* promotes those goals. So, we've got the first part of what we need – a ψ-ing, viz. acting to acquire no resources at all, such that the corresponding instance of (2*) is true. Now consider the corresponding instance of (3):

> (3-None) So: Almost whatever a superintelligent agent's goals, there is an instrumental reason for that agent to *a*ct to acquire extreme levels of physical resources *rather than act to acquire no resources at all.*

(3-None) is not dangerous. This is true even if it's dangerous for a superintelligent agent to actually act to acquire extreme levels of physical resources. This is because (3-None) says that superintelligent agents (almost always) have a particular contrastive instrumental reason, viz. a reason to act to acquire extreme levels of physical resources *rather than act to acquire no resources at all* no matter what their goals. But even if true, this means that superintelligent agents are (almost always) only dangerous in very specific circumstances, namely those situations in which they are faced only with the choice between acting to acquire extreme levels of physical resources and acting to acquire no resources at all.[45] Those are the cases where the instrumental reason will, if the agent is rational, incentivize them to behave in a dangerous way (say, by acting to acquire extreme levels of resources). But as with Brett's case, this is clearly not existentially worrying in the way Bostrom intends: remember, Bostrom's claim is that superintelligent agents will have reason to (e.g.) acquire extreme levels of resources in a "wide range of situations" (2014, p. 232); but the situation where one is faced with the forced choice between acting to acquire extreme levels of resources and no resources at all is not a "wide range" of situations, it's a single, *very specific* situation.

Moreover, we can easily imagine that superintelligent agents will rarely, if ever face such situations.[46] We could even arrange this ourselves by making it very easy for any agent to acquire *moderate* levels of resources. For example, one way to do this would be by ensuring that every agent – artificial, superintelligent, or human

[45] And we might add: there are no other more weighty (instrumental) reasons at stake. It's worth pausing over this point. In general, Bostrom's argument is made independently of any claims about the weight of the reasons agents might have (see also fn. 29, above). But even if it turns out that in every case agents will have instrumental reason to phi, it doesn't follow that we should expect them to regularly ϕ if the majority of situations are ones in which there is sufficiently strong reason to do something other than ϕ. I think it's plausible that even in "dangerous" situations where agents might have instrumental reason to acquire extreme levels of resources they will also have strong instrumental reasons to acquire moderate levels of resources, too. One important question left open by this argument is therefore what we should expect the balance of reasons to support. I discuss these issues in [removed for blind review].

[46] Importantly, this is true even if it's true that almost always when agents face that situation they have the instrumental reason to do something dangerous. In other words, the point here is not that it's false that almost always agents could face situations where they'd have the relevant instrumental reason (though see below on whether we can expect this to be true); instead the point is that even if they do almost always have the relevant contrastive instrumental reason, it doesn't follow that their having this contrastive instrumental reason is dangerous because we shouldn't expect that reason to (almost always) lead to the relevant dangerous behavior. Thanks to an anonymous referee for encouraging clarity on this point.

– has access to a universal basic income that guarantees access to moderate levels of resources. Then, the fact that there's a reason to acquire extreme levels of resources *rather than no resources at all* doesn't tell us anything at all about what even a "superintelligent" instrumentally rational agent will do, let alone what they will do, as Bostrom puts it, in "many situations".[47] But surely if superintelligent agents are existentially risky in the way that Bostrom intends we shouldn't be able to preclude them from posing a risk of harm to us by the relatively straightforward method of guaranteeing them a basic income!

The point here isn't that in order to avoid existential catastrophe we require a UBI; instead, the point is that Bostrom's argument that we faced a risk of existential catastrophe no matter what superintelligent agents might want does not deliver on its promise. The argument promised to show that agents would have instrumental reason to do dangerous things (e.g. acquire extreme levels of resources) no matter what they want, and in a wide range of circumstances. But it turns out that even if agents have an instrumental reason to acquire extreme levels of resources no matter what they want in a wide range of circumstances, it is only dangerous for them to have it in a relatively restricted range of circumstances. This is because the reason is not a reason to acquire extreme levels of resources *full stop*. Instead, the reason to acquire extreme levels of resources is, like other reasons, contrastive: it is a reason to do an extreme, dangerous thing *rather than a single specific safe thing*, and it is only dangerous for agents to have such a reason if we expect that the world is such that they will regularly be faced with making a rational choice between the extreme dangerous behavior and the very specific safe thing. But why would this be true?

More generally, in order for instances of (3*) to be existentially dreadful, it must be the case that the agent's having an instrumental reason to *do one thing rather than another thing* is itself existentially dreadful. It's certainly true that there are instances of (3*) where that's true. To that end, consider:

> (3-Some) Almost whatever a superintelligent agent's goals, there is an instrumental reason for that agent to *act to acquire extreme levels of physical resources rather than moderate levels of resources.*

Unlike (3-None), (3-Some) *is* a (contrastive) instrumental reason that it is intuitively dangerous for agents to possess in a wide range of circumstances: after all, a superintelligent agent that has instrumental reason to acquire extreme levels of physical resources *rather than moderate levels* is dangerous under the assumption that only moderate levels of resource acquisition (and not extreme ones) are compatible with human safety. But in order to arrive at (3-Some) as a dangerous conclusion of Bostrom's argument, we need to get there via the corresponding version of Dangerous Contrastive Convergent Promotion::

> (2-Some) Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources *rather than moderate levels of resources* promotes those goals.

[47] (Bostrom, 2012, 84).

Happily for humanity, (2-Some) is simply false. It simply isn't true that, (almost) *whatever one's goals* acting to acquire extreme levels of physical resources *rather than moderate levels of those resources* promotes those goals. Instead, this is only true if one's goals require extreme, rather than moderate, levels of resources. But there are many goals for which this is not in fact true.[48] Instead, many goals are equally as likely to be achieved given moderate levels of resources as they are given extreme ones.[49] For instance, consider my goal of buying a stick of gum right now at the corner store. If I have $1, this will be just as likely to be achieved as if I had $1MM.

This is an important point, so it's worth emphasizing. The intuition that *acting to acquire more resources always promotes one's goals* is very tempting, but we should resist it. It is true that acting to acquire more physical resources in some sense entails acting to acquire more "all-purpose means" to achieving one's goals. That intuition, and the attendant claims about instrumental reasons, is what we saw driving Bostrom's argument when it was working at a merely intuitive level (see § 1 for discussion). But we're now in a position to evaluate that claim in more careful detail. In detail, the claim should be understood as the claim that for any arbitrary goal the probability of that goal's being achieved given *extreme* levels of physical resources is greater than the probability of that goal's being achieved given *moderate* levels of physical resources. But as I'm now busy pointing out, that claim is clearly false: the goal of buying a stick of gum from the corner store right now is a case in point. The case of a goal that *precludes* possession of extreme levels of physical resources, e.g., the goal of being a member of the middle-class, is another. It's trivial to develop more examples.

One response to this might be that among all the goals that an agent might have, including ones that manifestly do not require (or may even preclude) acting to acquire extreme levels of physical resources, it turns out that a very large proportion of the goals that we can expect a superintelligent agent to have are such that they do in fact require acting to acquire extreme levels of resources rather than acting to acquire a moderate amount of resources in order to promote them.[50] There are two problems with this response. The first is that it is unclear what reason we would have for thinking that this is true. Why would we expect a large proportion of a superintelligent agent's goals to be such that acting to acquire extreme levels of resources rather than acting to acquire a moderate level of resources is necessary to promote

[48] In fact, there may be *infinitely* many goals for which this is not true. For discussion, see [removed for blind review]. Thanks to [removed for blind review] for this suggestion. Here I do not need this stronger claim. What matters is just that there are a wide range of very plausible goals such that their achievement is equally likely given moderate, as compared to extreme, levels of physical resources.

[49] Similar points are made by (Ngo, Chan, and Mindermann 2022; Grace, 2022; Drexler, 2019). For a formal analysis of the assumptions required in order to deliver the result that agents' goals will (almost) always require extreme levels of resources, see (Gallow, 2024). Thanks to an anonymous referee for pointers to the unpublished work on this topic.

[50] Thanks to an anonymous referee for suggesting this point.

them? We've been given no reason to think this is true, and it's not true that (e.g.) randomly sampling from among all possible goals delivers this result.[51]

The second, more serious problem with this reply is that it is not compatible with Bostrom's argument. Recall, Bostrom frames his argument as one that remains agnostic concerning the specific content of an agent's goals and argues from this agnosticism to a surprising, scary conclusion about what that agent has instrumental reason to do. What we've just seen is that, although a *simple* probabilistic account of promotion might get you from agnosticism to the scary conclusion, you can't get from agnosticism to that scary conclusion via contrastivism about promotion. Again, this is because the plausible cases of contrastive promotion, such as (2-None), do not entail the existence of scary contrastive instrumental reasons – (3-None) isn't worrying. And the scary cases of contrastive instrumental reasons, such as (3-Some), are not entailed by plausible cases of contrastive promotion – (3-Some) is false for a very wide range of goals, and so an argument that purports to remain agnostic on goals' content cannot appeal to it.

So, if we go contrastive about promotion, as it seems we must in order to make a *probabilistic* account of promotion work (§ 3.3), then the intuitive case for Dangerous (*Contrastive*) Convergent Promotion is undermined. Indeed, once we go contrastive, we can see that, rather than facing convergent instrumental reasons *regardless* of what goals superintelligent agents might have, the question of the instrumental reasons of such agents – grounded in facts about promotion – is extraordinarily sensitive to the content of their goals. Hence we do not get dangerous "convergent" instrumental reasons because we do not get dangerous convergent (contrastive) promotion.

4 Dangerous convergent fit-based promotion?

We just saw that a *probabilistic* account of promotion fails to deliver Dangerous Convergent Promotion. Is there a non-probabilistic account of promotion that would do the work required? The only systematic non-probabilistic account of promotion in the literature is the so-called "fit-based" account of promotion.[52] In this section, I first motivate and then lay out the details of the fit-based account of promotion.

[51] For more on this point see Gallow (forthcoming).

[52] The fit-based account of promotion has been developed both as a non-contrastive, non-probabilistic view that can be disjunctively paired with a probabilistic account of promotion (N. Sharadin, 2015a; 2016) and as a unified contrastive, probabilistic stand-alone view about promotion (Sharadin & Dellsén, 2019). For criticism of the non-contrastive, non-probabilistic fit-based account, see (Behrends and DiPaolo 2016; Elson, 2019; Lin, 2018). The contrastive, probabilistic stand-alone version of the fit-based view is able to capture the insights of (purely) probabilistic, i.e., non-fit-based accounts of promotion, and is thereby strictly superior. But here, I focus on the fit-based account in its non-contrastive, non-probabilistic form, since the criticisms of the version of Bostrom's argument that relies on a contrastive, probabilistic account of promotion developed above in § 3.4 apply equally well to a contrastive, probabilistically-inflected but fit-based account of promotion. For detailed discussion of how the fit-based account can be developed to take account of the insights of probabilistic accounts of promotion, see (Sharadin & Dellsén 2019).

Then, I consider whether a fit-based account of promotion can be used to support Dangerous Convergent (Fit-based) Promotion, and whether a version of Bostrom's argument that uses this premise is acceptable. To anticipate: it cannot, and it is not.

4.1 Fit-based promotion

The intuitive idea behind a fit-based account of promotion is easy to grasp. Goals are distinguished from (e.g.) beliefs in terms of their "direction of fit": whereas beliefs aim to to fit the world, goals aim to make world fit them.[53] When the content of the world (perfectly) matches the content of a goal, we can say that the goal (perfectly) "fits" the world. So understood, fit comes in degrees. For instance, consider Sia's goal of counting all the grains of sand on a beach. A world where Sia has counted 99% of the grains of sand enjoys a better "fit" with her goal of counting all the grains of sand than a world where Sia has counted only 50% of the grains. An action's *promoting* a goal, then, can be understood (roughly) as a matter of the action's leading to an increase in the degree of fit between the goal and the world. Intuitively: promoting a goal is a matter of making the world more like the goal represents it as being (when it's achieved). For instance, each time Sia counts another grain of sand her action *promotes* her goal of counting *all* the grains of sand by increasing the degree of fit between her goal and the world.

More carefully, we can define an order $\succsim_G$ on the set of possible worlds W so that for any goal G and for any two worlds w_1 and w_2, $w_1 \succsim_G w_2$ just in case w_1 fits G at least as well as w_2; and then we have it that w_1 is a *strictly better* fit for G than w_2, i.e.., $w_1 \succ_G w_2$, just in case $w_1 \succsim_G w_2$ but not $w_2 \succsim_G w_1$.[54] Using this ordering of worlds in terms of their fit with goals, where w_φ refers to the world where the agent φs and $w_@$ refers to the actual world prior to the agent's φ-ing, we have:

Fit-Promote: Agent A's φ-ing promotes goal G iff $w_\varphi \succ_G w_@$.

Intuitively, in Sia's case, let φ-ing be counting an additional grain of sand, and let G be counting all the grains of sand. Then, Sia's φ-ing promotes her goal, since the world where she counts an additional grain of sand (w_φ) enjoys strictly better fit with her goal of counting all the grains of sand than the actual world.[55]

Notice that this fit-based account of promotion does not require that an action increase the *probability* of a goal's achievement in order to promote it. According to its proponents, this is part of its appeal.[56] For, there are some cases where a goal intuitively can be promoted but where it's strictly impossible to increase the probability of that goal's achievement. Goals that are impossible to achieve are like this.

[53] (Sharadin, 2015a). The classic source for this idea is (Anscombe 1957). Here, nothing I say relies on making good on the idea of "direction of fit" in any robust way. For recent criticism of the "very idea" of direction of fit, see (Frost, 2014).

[54] For completeness: w_1 and w_2 fit G *equally well*, $w_1 \sim_G w_2$, just in case $w_1 \succsim_G w_2$ and $w_2 \succsim_G w_1$. See (Sharadin, 2015a) for discussion.

[55] Notice that in order to say this we do not need to settle on a single articulable account of similarity or "fit" anymore than we need to do the equivalent thing in order to make use of the idea of similarity between possible worlds in our semantics. See (Lewis, 1973).

[56] (N. Sharadin 2015a; 2016; N. Sharadin and Dellsén 2019).

For instance, consider the goal of achieving nothing. It is impossible to achieve this goal. If the goal is achieved, then it is not achieved. But if it is not achieved, then it is also not achieved. So, this goal cannot be achieved: the probability of its achievement is *always* zero. Probabilistic accounts of promotion are therefore required to say that such a goal cannot be promoted.[57]

Nevertheless, intuitively at least, it *is* possible to promote the achievement of the goal of achieving nothing. For instance, suppose an agent with that goal is given an opportunity to *frustrate* the achievement of one of her *other* goals by φ-ing. Then, φ-ing intuitively promotes the goal of achieving nothing: by frustrating the achievement of one of their other goals, the agent would "move closer" to a world where they achieve nothing. A fit-based account of promotion can easily account for this fact, since a world that's the result of φ-ing is one that better fits the goal of achieving nothing (since it's a world where less is achieved, and achieving less is fit-wise closer to achieving nothing).

In any case, my aim here isn't to argue in favor of a fit-based account of promotion. Instead, I aim to investigate whether a fit-based account of promotion can do the work required by Bostrom's argument. I'll now argue that it cannot.

4.2 Dangerous convergent fit-based promotion?

Recall, we are interested in:

> **Dangerous Convergent Promotion (DCP):** Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources promotes those goals.

According to the present account, what it is for an action to "promote" a goal is for the action to increase the degree of fit between the world and the goal. Therefore, DCP should be understood as claiming that (almost) whatever the content of a superintelligent agent's goals, the degree of fit between the actual world and the agent's goals is greater if they act to acquire extreme levels of physical resources, i.e.,

> **Dangerous Convergent Fit-based Promotion (DCFPB):** Almost whatever a superintelligent agent's goals G, $w_{\text{act to acquire extreme levels of resources}} \succ_G w_@$.

Whether DCFBP is true will turn on whether (almost) all of a superintelligent agent's goals are such that worlds where an agent acts to acquire extreme levels of resources are a strictly better fit than worlds where the agent doesn't so act. We have been given no reason to think this is true. Recall, fit is a matter of match between the content of a goal and the content of the world – the greater the degree of match

[57] Proponents of a fit-based view of promotion (e.g. (Sharadin, 2015a; 2016; Sharadin & Dellsén 2019)) therefore appeal to these kinds of cases in arguing against various probabilistic accounts of promotion. Here, I remain neutral on whether a fit-based account of promotion is correct: my aim is to investigate whether such an account can do the work required in Bostrom's argument in delivering Dangerous Convergent Promotion.

between the content of a goal and the content of the world, the greater the degree of fit. But so long as a goal's content doesn't itself involve the acquisition of extreme levels of physical resources, a world where an agent acts to acquire extreme levels of physical resources is not a world that is a better match, a strictly better fit, with that goal.

Consider by way of illustration one of the goals Bostrom himself mentions, and that we used in illustrating the view above: the goal of counting grains of sand. Now consider the ordering of possible worlds $W_{low} \ldots W_{extreme}$ induced by sorting worlds in terms of the total amount of resources an agent has, where W_{low} is the world (or worlds) where the agent has the lowest possible amount of physical resources and $W_{extreme}$ is the world (or worlds) where the greatest possible amount of physical resources. Even if we assume that $W_{\text{act to acquire extreme levels of resources}}$ is among $W_{extreme}$,i.e., a world where the agent acts to acquire extreme levels of physical resources is one where the agent *actually acquires them*, we still lack a reason to think that for any arbitrary goal G, $W_{extreme} \succ_G W_{low}$. In other words: Why think the content of any arbitrary goal is a better match for worlds where an agent has extreme, rather than low, moderate, or *anywhere in between*, levels of resources? In general, this will not be true unless the content of a goal *itself implicates extreme levels of physical resources*. Of course, if a superintelligent agent's goal involves acting to acquire extreme levels of physical resources (say, as a necessary means to achieving the goal), then a world where they do so will in fact be a better fit for their goal.

But Bostrom and others who go in for instrumental convergence arguments are in no position to assume that the achievement of a superintelligent agent's arbitrary goals will necessarily involve the acquisition of extreme levels of physical resources. That would vitiate the purpose of such arguments! Remember, the purpose of Bostrom's argument is to convince us that we have reason to be afraid of (the instrumental reasons of) superintelligent agents (almost) *whatever the content of their goals*, i.e., even if their goals do not themselves involve acting to acquire extreme levels of physical resources. That is why Bostrom's argument is about the existence of dangerous *promotion* (and so dangerous *instrumental reasons*), and not simply the existence of dangerous *goals*! But if what it is to promote a goal (and so what it take for there to be an instrumental reason) is to increase the *fit* between the world and the goal, then as we can now see unless the goal itself involves dangerous content (e.g. the acquisition of extreme levels of physical resources, or power, or whatever), then promoting the goal by way of moving to a world with better *fit* between the goal and the world won't itself be dangerous.

It's worth noting that a world where an agent acts to acquire extreme levels of resources could of course make it *more likely* that they will achieve (e.g.) a complete count of grains of sand (say by giving them time to work on achieving their insane goal). (This is true even if *acquiring* extreme levels of resources does not increase the degree of fit between the world and that goal.) One response on Bostrom's behalf, then, is to say that increasing the likelihood of a goal's being achieved

is (perhaps *also*) what promoting a goal comprises.[58] But if we go this route, we are back to the problems with a probabilistic account of promotion: we shall need to pick a baseline (§ 3.2), this motivates going contrastive (§ 3.3), and as we've already seen, having gone contrastive Bostrom's argument loses its sting (§ 3.4).

Here is another way to put the problem with attempting to use a fit-based account of promotion in argument in favor of Dangerous Convergent Promotion. According to a fit-based account of promotion, an action promotes a goal when the fit between the goal and the world goes up. But many, perhaps most times, the fit between a goal and the world is completely insensitive to a range of changes in the total amount of resources an agent has. Consider again my goal of buying a stick of gum. Suppose I have a million dollars. Now suppose I can φ. By φ-ing I bring about a world where I have a million and one dollars. Is there any instrumental reason at all (however small) for me to φ given my goal of buying gum? That depends on whether φ-ing *promotes* my goal. According to a fit-based account of promotion, φ-ing promotes my goal just in case a world where I have a million plus one dollars better fits my goal of buying a stick of gum than a world where I have a million dollars fits my goal of buying a stick of gum. But it doesn't. From the point of view of my gum goal, the actual world (where I have a million dollars) and the φ world (where I have a million plus one dollars) manifest the same fit; this is because nothing in my goal requires having more dollars. Moreover, the same goes for a very wide range of other kinds of physical resources, other goals, etc..[59]

The result is that, given a fit-based account of promotion, it simply isn't true that various dangerous actions, such as acting to acquire extreme levels of physical resources, promote superintelligent agent's goals, almost whatever those goals' content. Instead, whether this is so depends essentially on exactly what the content of those goals is. In other words, we should reject Dangerous Convergent Promotion. This blocks argument to the worrying conclusion concerning superintelligent agent's Dangerous Convergent Instrumental Reasons.

5 Open-ended goals?

The scary conclusion of Bostrom's argument is.

[58] Proponents of fit-based accounts of promotion are themselves sensitive to the need to allow that increasing the probability of a goal's achievement is a way of promoting it. Hence a fit-based account of promotion is (typically) paired with a probabilistic account in one way or another, either by disjoining it or by probabilizing the relevant fit-based account directly. For examples of each of these approaches, see (Sharadin, 2015a) and (Sharadin & Dellsén 2019), respectively.

[59] Perhaps you think it's *always* possible to induce a difference in "fit" between worlds and goals, because worlds will always be dissimilar in some respect. But this isn't so. Imagine a world where you are Treasury Secretary and married to the President of the United States. Is that a world that is more or less similar to the world where you are Treasury Secretary and *engaged* to the President of the United States? Which world is more similar to the actual world? These are bad questions. For discussion, see (Lewis, 1973).

Dangerous Convergent Instrumental Reason: Almost whatever a superintelligent agent's goals, there is an instrumental reason for that agent to act to acquire extreme levels of physical resources.

This conclusion relies, as we've seen, on a claim about promotion, namely.

Dangerous Convergent Promotion: Almost whatever a superintelligent agent's goals, acting to acquire extreme levels of physical resources promotes those goals.

I've just argued that on two extant, plausible accounts of what it is for an action to promote a goal – the contrastive probabilistic and the fit-based accounts, this claim about promotion is false. In my discussion I appealed to two kinds of goals in order to illustrate why Dangerous Convergent Promotion is false. First, I appealed to goals that are very *easy* to achieve (such as buying a stick of gum) where extreme levels of physical resources do not either contrastively or fit-wise promote a goal. Second, I appealed to goals that are *humble* (such as the goal of being a member of the economic middle-class), where acting to acquire extreme levels of physical resources in fact (contrastively or fit-wise) *dis*promotes the relevant goal.[60]

In his argument for Dangerous Convergent Instrumental Reason, Bostrom largely ignores humble and easy goals. Instead, he appeals to what we can call *open-ended* goals, such as the goal of "calculating the decimals of pi" (Bostrom, 2014, 84). So, it might seem that my argument, which focuses on easy and humble goals, misses the mark. After all, for open-ended goals, it can *still* intuitively seem like (e.g.) acting to acquire extreme levels of physical resources will (contrastively or fit-wise) promote those goals. Couldn't an agent with the goal of calculating the decimals of pi always put more (and more, and more) physical resources to use? Perhaps they could use those resources to build increasingly large datacenters for storing the calculated digits, or to research superconductors in order to improve efficiency at the level of compute. Isn't it just obvious, at least when it comes to open-ended goals, that certain kinds of dangerous activities will be ones that agents have reason to do? This seems to be the kind of thought Bostrom has in mind.[61] And I admit that it can seem very intuitive that this is true, especially when thinking about open-ended goals; *more* just seems like it must be *always* better! So, again, it might seem as if my argument misses the mark. I have two replies to this line of thought.

The first reply is partly conciliatory. Suppose open-ended goals (such as calculating the digits of pi) really are such that, at least in general, doing dangerous things such as acting to acquire extreme levels of physical resources in fact promotes those goals. (Below, I'll again explain why we shouldn't in fact accept this claim.) Again, this fact about open-ended goals stands in contrast to the facts about easy and humble goals, where as we've already seen, it isn't true that (e.g.) acting to acquire extreme levels of physical resources promotes those goals. In any case, if it is true

[60] Thanks to a referee for suggesting this way of putting the point, and for encouraging the discussion of open-ended goals in this section.

[61] See (Bostrom, 2012), esp. the discussion in fn. 18, p. 82.

with respect to a goal with open-ended content that acting to acquire extreme levels of physical resources promotes such a goal, then a superintelligent agent's having such an open-ended goal would indeed be scary. But in that case, in order for Bostrom to establish his scary conclusion (which, recall, is about superintelligent agents in general), he would have to establish a claim about the relative proportion of open-ended goals to non-open-ended (e.g., easy or humble) goals among superintelligent agents in general. One way to do this would be to argue that open-ended goals comprise most of the space of possible goals. But it's hard to see why this would be so, and in any case Bostrom hasn't mounted a defense of this implausible claim.[62] So, even if it's true that doing dangerous things in general promotes open-ended goals (again, I'm about to reiterate my view that this is not true), this is neither here nor there with respect to whether superintelligent agents will have reason to do dangerous things: that will turn on whether they (are likely to) have open-ended goals.

The second reply relies on accepting a contrastive account of promotion such as the one articulated above (Sect. 3.3). Recall, the crucial insight of the contrastive account of promotion is that an agent's action never simply promotes (or dispromotes) a goal. Instead, an agent's action *rather than some other action* promotes (or dispromotes) a goal. But now notice that, even when it comes to open-ended goals for which in general the *possession* of extreme levels of physical resources as compared to the *possession* of moderate levels of physical resources would be useful, it's unclear why we would think that in general *acting to acquire* extreme levels of physical resources as compared to *acting to acquire* moderate levels of physical resources makes the satisfaction of that goal more likely. To explain: the distinctive mark of open-ended goals (we're here assuming) is that, because they are open-ended, the possession of extreme levels of physical resources will be more useful in pursuing their satisfaction than the possession of (merely) moderate levels of physical resources. But admitting that having and being able to use extreme levels of physical resources is in some sense generally better with respect to the satisfaction of a goal than having and being able to use (merely) moderate levels of physical resources is not the same as admitting that in general *pursuing* extreme levels of physical resources is more useful with respect to that goal than *pursuing* moderate levels of physical resources! More prosaically: although I'd happily *possess* many millions of dollars (that I could then deploy to promote my goals), I'd rather not *pursue* those millions.

Here's another way to put this point. Compared to acting to acquire moderate levels of physical resources, acting to acquire extreme levels of physical resources is typically more difficult, more logistically challenging, and has a very particular set of possible failure modes, some of which could easily be more devastating than any way of failing to acquire moderate levels of physical resources.[63] So, even if having extreme levels of resources generally makes it easier to achieve many (especially open-ended) goals, it doesn't follow that acting to acquire extreme levels of physical resources generally increases the likelihood that a goal will be satisfied as compared to acting to acquire moderate levels of physical resources. Hence the admittedly

[62] Moreover, as (Gallow, 2024) has argued, we have positive reason to think that this claim is false.

[63] Thanks to an anonymous referee for this way of putting the point.

intuitive idea that, especially when it comes to open-ended goals, "more (resources) is better" doesn't have any probative value with respect to whether, including when it comes to open-ended goals, the *pursuit* of more is better as compared to available alternatives, in particular the alternative of pursuing (somewhat) less.

6 Concluding remarks

Instrumental convergence arguments have a scary conclusion: superintelligent agents may have instrumental reason to do dangerous things regardless of their specific goals. If this conclusion is true, it is very worrying. But establishing this conclusion relies on having an account of what it takes for dangerous, scary things to "promote" an agent's goals.

There are two extant accounts of promotion in the philosophical literature: probabilistic accounts and fit-based accounts. Neither account of promotion supports the scary conclusion. Instead, both accounts of promotion entail that whether agents will have instrumental reason to do dangerous, scary things depends essentially on the specific content of their goals. And, of course, we should all agree that *if* agents have certain certain goals, then this could be very dangerous indeed! That is simply what follows from facts about the nature of instrumental rationality and the nature of specific goals.

But we should reject the idea that agents have these scary reasons almost *whatever it turns out they want*, or (almost) whatever the specific content of their goals. Hence you can't get scary results from combining the orthogonality thesis, which says that agents could in principle have any goals whatsoever, with facts about what kinds of behavior would promote *some particular subset of the goals they could have*. It doesn't help to shift the focus to open-ended goals, since the results in that case are largely the same; for one thing, we haven't yet been given any reason to suppose that (potentially) dangerous open-ended goals are very likely, and for another thing whether or not a pursuit of extreme levels of physical resources will actually promote an agent's open-ended goals strongly depends on the alternatives available. The result is that, instrumental convergence arguments do not support the idea that, as Bostrom puts it, the "default outcome of the creation of machine superintelligence is existential catastrophe".[64]

Acknowledgements This paper was largely written during my term as a Philosophy Fellow at CAIS. Thanks to Mitch Barrington, Cameron Domenico Kirk-Giannini, William D'Alessandro, Frank Hong, Simon Goldstein, Jacqueline Harding, Nick Laskowski, Harry Lloyd, Robert Long, and Elliot Thornley for helpful feedback and discussion. Thanks also to three anonymous referees for this journal for helpful feedback that improved (but lengthened) the paper.

Funding Research Grants Council, University Grants Committee, 17602622, Nathaniel Paul Sharadin

[64] (Bostrom, 2014, 140).

References

Anscombe, Gertrude Elizabeth Margaret. (1957). *Intention*. Vol. 13. Cornell University Press.

Bales, A., D'Alessandro, W., & Kirk-Giannini, C. D. (2024). Artificial intelligence: Arguments for catastrophic risk. *Philosophy Compass, 19*(2), e12964. https://doi.org/10.1111/phc3.12964

Behrends, J., & DiPaolo, J. (2011). Finlay and Schroeder on promoting a desire. *Journal of Ethics and Social Philosophy, 6*(1), 1–7. https://doi.org/10.26556/jesp.v6i1.146

Bales, A., D'Alessandro, W., & Kirk-Giannini, C. D. (2016). Probabilistic promotion revisited. *Philosophical Studies, 173*(7), 1735–1754. https://doi.org/10.1007/s11098-015-0576-0

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines, 22*(2), 71–85. https://doi.org/10.1007/s11023-012-9281-3

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

Broome, J. (2013). *Rationality through reasoning*. Blackwell: Wiley.

Carlsmith, Joseph. (2022). Is power-seeking AI an existential risk? arXiv preprint arXiv, https://doi.org/10.48550/arXiv.2206.13353.

Christian, B. (2020). *The alignment problem: Machine learning and human values*. New York: Norton.

Coates, D. J. (2013). An actual-sequence Theory of promotion. *Journal of Ethics and Social Philosophy, 7*(3), 1–8. https://doi.org/10.26556/jesp.v7i3.157

Côté-Bouchard, C. (2015). Epistemic instrumentalism and the too few reasons objection. *International Journal of Philosophical Studies, 23*(3), 337–355. https://doi.org/10.1080/09672559.2015.1042007

Cowie, C. (2014). In defence of instrumentalism about epistemic normativity. *Synthese, 191*(16), 4003–4017. https://doi.org/10.1007/s11229-014-0510-6

Dorsey, D. (2012). Subjectivism without desire. *Philosophical Review, 121*(3), 407–442. https://doi.org/10.1215/00318108-1574436

Dorsey, D. (2021). *A theory of prudence*. Oxford: Oxford University Press,.

Drexler, K. Eric. (2019). Reframing superintelligence comprehensive AI services as general intelligence. Technical Report 2019–1. Future of Humanity Institute, University of Oxford.

Dyke, M. M. (2021). Could our epistemic reasons be collective practical reasons? *Noûs, 55*, 842–862. https://doi.org/10.1111/nous.12335

Elson, Luke. (2019). Probabilistic promotion and ability. *Ergo: An Open Access Journal of Philosophy,* https://doi.org/10.3998/ergo.12405314.0006.034.

Evers, D. (2009). Humean agent-neutral reasons? *Philosophical Explorations, 12*(1), 55–67. https://doi.org/10.1080/13869790802635614

Finlay, S. (2006). The reasons that matter. *Australasian Journal of Philosophy, 84*(1), 1–20. https://doi.org/10.1080/00048400600571661

Finlay, S. (2014). *Confusion of Tongues: A theory of normative language*. Usa: Oxford University Press.

Frost, K. (2014). On the very idea of direction of fit. *Philosophical Review, 123*(4), 429–484.

Gallow, J. Dmitri. (2024). Instrumental Divergence. *Philosophical Studies*, https://philarchive.org/rec/GALIDB.

Grace, Katja. (2022). Counterarguments to the Basic AI Risk Case. Substack newsletter. *World Spirit Sock Stack* (blog), October 14, 2022. https://worldspiritsockpuppet.substack.com/p/counterarguments-to-the-basic-ai.

Grice, H. Paul. (1975). Logic and Conversation. In *The Semantics-Pragmatics Boundary in Philosophy*, Maite Ezcurdia and Robert J. Stainton. (Eds.), 47. Broadview Press.

Hannon, Michael, and Elise Woodard. The Construction of Epistemic Normativity.

Hendrycks, Dan. (2023). Natural Selection Favors AIs over Humans. arXiv.Org. March 28, 2023. https://arxiv.org/abs/2303.16200v3.

Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside. (2023). An Overview of Catastrophic AI Risks. *arXiv Preprint* arXiv:2306.12001.

Kant, Immanuel. (1785). *Groundwork for the Metaphysics of Morals*. Oxford University Press.

Kornblith, H. (1993). Epistemic Normativity. *Synthese, 94*(3), 357–376. https://doi.org/10.1007/bf01064485

Korsgaard, C. M. (1983). Two distinctions in goodness. *Philosophical Review, 92*(2), 169–195. https://doi.org/10.2307/2184924
Korsgaard, Ch. (1996). *The sources of normativity*. Cambridge University Press.
Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford University Press.
Lewis, D. K. (1973). *Counterfactuals*. Blackwell.
Lin, E. (2018). Simple probabilistic promotion. *Philosophy and Phenomenological Research, 96*(2), 360–379. https://doi.org/10.1111/phpr.12310
Mitchell, Melanie. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin UK.
Ngo, Richard, Lawrence Chan, and Sören Mindermann. (2022). The alignment problem from a deep learning perspective. https://arxiv.org/abs/2209.00626v5.
Omohundro, Stephen M. (2007). Self-Aware Systems. n.d. The Nature of Self-Improving Artificial Intelligence.
Rawls, John. (1971). *A Theory of Justice: Original Edition*. Belknap Press.
Rawls. (1999). *A Theory of Justice: Revised Edition*. Harvard University Press.
Rawls. (2000). *Lectures on the History of Moral Philosophy*. Barbara Herman (Ed.), Cambridge, Mass.: Harvard University Press.
Rawls. (2001). *Justice as Fairness: A Restatement*. Harvard University Press.
Rinard, S. (2015). Against the new evidentialists. *Philosophical Issues, 25*(1), 208–223. https://doi.org/10.1111/phis.12061
Rinard, S. (2017). No exception for belief. *Philosophy and Phenomenological Research, 94*(1), 121–143. https://doi.org/10.1111/phpr.12229
Rinard, S. (2019). Equal treatment for belief. *Philosophical Studies, 176*(7), 1923–1950. https://doi.org/10.1007/s11098-018-1104-9
Schroeder, M. (2007). *Slaves of the Passions*. Oxford University Press.
Sharadin, N. (2015a). Problems for pure probabilism about promotion (and a Disjunctive Alternative). *Philosophical Studies, 172*(5), 1371–1386. https://doi.org/10.1007/s11098-014-0354-4
Sharadin, N. (2015b). Reasons and promotion. *Philosophical Issues, 25*(1), 98–122. https://doi.org/10.1111/phis.12057
Rawls. (2016). Checking the neighborhood: A Reply to DiPaolo & Behrends on Promotion. *Journal of Ethics and Social Philosophy, 10*(1), 1–8. https://doi.org/10.26556/jesp.v10i1.181
Sharadin, N. (2018). Epistemic Instrumentalism and the Reason to Believe in Accord with the Evidence. *Synthese, 195*(9), 3791–3809. https://doi.org/10.1007/s11229-016-1245-3
Sharadin, N. (2019). Ecumenical Epistemic Instrumentalism. *Synthese, 198*(3), 2613–2639. https://doi.org/10.1007/s11229-019-02232-7
Rawls. (2024). How Strong Are The Instrumental Reasons to Destroy Humanity?
Sharadin, N., & Dellsén, F. (2019). Promotion as Contrastive Increase in Expected Fit. *Philosophical Studies, 176*(5), 1263–1290. https://doi.org/10.1007/s11098-018-1062-2
Sharadin, N. P. (2022). *Epistemic Instrumentalism Explained*. Routledge.
Shulman, Carl. (2010) Omohundro's 'Basic AI Drives' and Catastrophic Risks.
Sinnott-Armstrong, W. (2008). A Contrastivist Manifesto. *Social Epistemology, 22*(3), 257–270. https://doi.org/10.1080/02691720802546120
Snedegar, J. (2014). Contrastive Reasons and Promotion. *Ethics, 125*(1), 39–63. https://doi.org/10.1086/677025
Snedegar, J. (2017). *Contrastive Reasons*. Oxford University Press.
Snedegar, J. (2019). Deliberation, Reasons, and Alternatives. *Pacific Philosophical Quarterly, 100*(3), 682–702. https://doi.org/10.1111/papq.12262
"The Basic AI Drives." (2007). *Self-Aware Systems* (blog). November 30, 2007. https://selfawaresystems.com/2007/11/30/paper-on-the-basic-ai-drives/.
Turner, Alexander Matt, Logan Riggs Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. (2021). Optimal Policies Tend To Seek Power. In . https://openreview.net/forum?id=l7-DBWawSZH.
Weatherson, Brian. 2006. "Questioning Contextualism." In *Aspects of Knowing: Epistemological Essays*, edited by Stephen Cade Hetherington, 133–47. Elsevier.
Whiting, D. (2023). Admiration, Appreciation, and Aesthetic Worth. *Australasian Journal of Philosophy, 101*(2), 375–389. https://doi.org/10.1080/00048402.2021.1986556