Superintelligent AI, Instrumental Convergence, and Promotion

It seems increasingly likely that we will someday develop human-level general artificial intelligence and that we will later witness the emergence of *artificial superintelligence (ASI)*—AI that is much more intelligent than human beings are across all domains. Although ASIs could take the form of 'oracles' that can do nothing but answer the questions they are posed[1], they could equally well take the form of agents—whether dispersed as software or embodied as robots—that deliberate, plan, and act in pursuit of their goals. Intelligence is positively correlated in agents with effectiveness in the pursuit of one's goals, so ASIs could well be extremely capable agents. Since their emergence is a serious possibility, we should try to infer what we can about what they would do. In this paper, I consider Nick Bostrom's "instrumental convergence" argument that ASIs are likely to attempt to do things that threaten the survival of humanity.[2] After summarizing the argument, I defend it from a recent objection due to Nathaniel Sharadin.[3] In brief, Bostrom's argument is that a wide range of possible ASIs with different final goals would have instrumental reasons to pursue, and would thus be likely to pursue, certain instrumental goals the pursuit of which would be dangerous to humanity. Sharadin argues that insights from the philosophical literature on what it takes for an action to promote the attainment of a goal undermine this argument by showing that we needn't accept its claims about the instrumental reasons of ASIs. I argue that they do no such thing. Although the argument may have other problems, the philosophical literature on promotion provides no successful objection to it.

Two more general lessons will emerge along the way. The first is that we shouldn't overstate what the instrumental convergence argument purports to show: on its most plausible interpretation, its conclusion isn't that ASIs would attempt to do things that threaten the survival of humanity but merely that this possibility is likely or plausible enough to merit our concern. The second is that ASIs' motivations may be dependent on their capabilities in a way that is underappreciated: if the argument is on the right track, then how likely an ASI is to aim or attempt to do things that are dangerous to humanity depends on how capable it is, and limiting its capabilities can make it less likely to aim or attempt to do such things.

[1] Bostrom (2014), pp. 177-81.
[2] Bostrom (2014), ch. 7. This argument was anticipated in Omohundro (2008).
[3] Sharadin (2025).

1. The Instrumental Convergence Argument

The instrumental convergence argument purports to establish only that ASIs are likely to *attempt* or *aim* to eliminate or otherwise greatly harm humanity. On its own, it cannot and does not purport to show that they would likely *succeed* in doing that: whether they would depends on how powerful or capable they are, which is beyond the scope of the argument to determine. But although the success of the argument would not, on its own, establish that the emergence of an ASI would likely result in great harm to humanity, it would make this conclusion plausible enough to merit concern. After all, an ASI could, in principle, take the form of an armed robot or command an army of them, hack into and take over our defense systems, or design and arrange for the synthesis and dispersion of a deadly pathogen.[4] Absent good reasons for thinking that ASIs would lack such capabilities, we can't assume that their attempts to eliminate or otherwise greatly harm us would fail.

Some terminology will be necessary to understand the argument. An *instrumental* goal is a goal that an agent has because it believes that its attainment would promote the attainment of one or more of its other goals. An *intrinsic* or *final* goal is a goal that an agent has, and not just because it believes that its attainment would promote the attainment of one or more of its other goals. For most people, for example, making money is an instrumental goal and being happy is a final goal. A goal can be both final and instrumental: someone who fetishizes money can want to make money both instrumentally and finally, and someone who wants to be well liked and who believes that happy people are better liked can want happiness both finally and instrumentally. When an agent's performing a particular action would promote (i.e., facilitate or be conducive to) the attainment of one of its goals, whether final or instrumental, this fact gives that agent an *instrumental reason* to perform that action. If one of your goals is to make money, for example, then the fact that your looking for a job would promote the attainment of this goal gives you an instrumental reason to look for a job.

It might seem that whether an ASI will attempt to act in ways that are dangerous to humanity turns on whether its final goals are dangerous to humanity. Since it is unlikely that an ASI will be created by a doomsday cult instead of a technology company or a governmental project, it might therefore seem unlikely that there will be an ASI that is disposed to be dangerous to humanity. The upshot of the instrumental convergence argument is that this natural line of thought is mistaken. Since "[ASIs]

[4] Bostrom (2014), ch. 6.

having any of a wide range of final goals will nevertheless pursue similar intermediary goals because they have common instrumental reasons to do so," and because their pursuit of these intermediary goals would likely endanger humanity, it is sufficiently likely or plausible—that is, likely or plausible enough to merit our concern—that ASIs would try to do things that endanger humanity.[5]

The argument starts with the contention that "[s]everal instrumental [goals]… are convergent in the sense that their attainment would increase the chances of [an ASI's] goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental [goals] are likely to be pursued by a broad spectrum of [ASIs]."[6] These *convergent instrumental goals* are those of self-preservation, retaining one's final goals, amplifying one's intelligence, improving one's technology, and acquiring more resources. Suppose, for example, that an ASI has the final goal that a particular set of unsolved problems in physics be solved. The attainment of that goal would be promoted by that ASI's continued existence, provided that it retains that goal: after all, the relevant problems are more likely to be solved if there is an ASI who intends to solve them.[7] It would be promoted by the ASI's amplifying its intelligence and improving its technology, since such enhancements would make it more capable of solving such problems. And it would be promoted by the ASI's engaging in an unlimited process of resource acquisition, since more resources would always yield more computing power with which to search for and verify solutions. The ASI's final goal would therefore give it instrumental reasons to act so as to ensure its continued existence, retain its final goals, amplify its intelligence, improve its technology, and acquire resources without limit. This reasoning generalizes to many other final goals: for a wide range of possible ASIs with different final goals, those ASIs would have instrumental reasons to pursue those convergent instrumental goals because this would promote the attainment of those final goals. Because any ASI that has these instrumental reasons would surely be intelligent enough to know that it has them, any such ASI would be likely to pursue those instrumental goals.

Next, the argument claims that an ASI that pursues those instrumental goals would likely attempt to eliminate or otherwise greatly harm humanity. After all, our continued existence at anything close to

[5] Bostrom (2014), p. 127.
[6] Bostrom (2014), p. 132.
[7] This is plausible even if the ASI's final goal is not that *it itself* solve those problems, which can't occur unless it exists, but simply that the problems *be solved*, which can occur even if it doesn't exist. Something can promote the attainment of a particular goal even if it isn't necessary for its attainment.

our current population and with anything like our current capabilities could threaten its attainment of those goals. Unless we are eliminated or greatly disempowered, we could try to destroy the ASI, alter its final goals, or limit or stop its cognitive enhancement, technological perfection, or resource acquisition. Indeed, we surely would try to limit its resource acquisition at least eventually, since we would need some resources to survive. For these reasons, the argument concludes, a wide range of possible ASIs with different final goals—even ones that seem perfectly innocuous—would be likely to act in ways that are dangerous to humanity. Thus, the probability that ASIs would try to do things that endanger humanity is sufficiently high to merit concern.

How exactly the argument should best be understood is somewhat unclear. For example, although intelligence is positively correlated with acting as one believes one has instrumental reasons to act, it's unclear how exactly we're supposed to get to conclusions about the likely behavior of an ASI from premises about its knowledge of its instrumental reasons. Because there are situations in which it can be perfectly rational or intelligent not to do something that one knows one has instrumental reason to do (e.g., because one knows that one has stronger reasons not to do it), it would be nice to know what assumptions justify these inferences. Furthermore, the connection between the pursuit of the convergent instrumental goals and the performance of actions that threaten the survival of humanity also merits clarification. How much work is being done by the claim that our continued existence would eventually limit an ASI's resource acquisition? Why not think that ASIs would, out of caution, pursue those goals in ways that are friendly to us? Bostrom's impressionistic statement of the argument doesn't answer these questions. However, since my aim is not to determine whether the argument succeeds but merely to evaluate Sharadin's objection to it, we won't need to answer them. That objection focuses entirely on the argument's first step, which concerns the instrumental reasons of a wide range of ASIs with a wide range of final goals.

One point must be emphasized, though. Although Bostrom sometimes proceeds as though "almost any final goal" would give every ASI that has it instrumental reason to pursue the aforementioned convergent instrumental goals[8], which suggests that he aims to show that ASIs would *almost certainly* try to act in ways that threaten the survival of humanity, he isn't most charitably read as making such strong claims. After all, he more often depicts his argument as merely concerning the instrumental

[8] Bostrom (2014), p. 132.

reasons of a "broad spectrum" of ASIs with a "wide range" of final goals.[9] Such an argument can't show that ASIs would almost certainly behave in a particular way: it can only show that, since it is sufficiently likely that ASIs, if they exist, would include ones in that broad spectrum, it is sufficiently likely that an ASI would behave in a particular way. Indeed, the argument isn't even most charitably construed as holding that a wide range of final goals would give *every* ASI that has them instrumental reasons to pursue the convergent instrumental goals. This is good news for the argument because even that weaker claim is, on reflection, too strong. After all, whether a final goal provides an agent that has it instrumental reasons to perform a particular action depends on the agent's capabilities and circumstances. For example, whether an agent with a resource-intensive final goal has instrumental reason to try to acquire more resources depends on whether it is able, in its present circumstances, to acquire more resources: if it isn't (e.g., because any of its attempts to do so would backfire), then its final goal gives it no instrumental reason to make such attempts. ASIs are potentially very capable agents, but they are not by definition so powerful that they will always be able to effectively pursue their goals no matter what circumstances they are in. Since some possible ASIs will, because of their limitations and circumstances, be incapable of making any progress in the pursuit of the convergent instrumental goals, it's not true that a wide range of final goals would give *every* ASI that has them instrumental reasons to pursue those goals. We should therefore read the argument as claiming only that a wide range of possible ASIs (namely, those with final goals within a particular wide range *and whose capabilities are sufficiently great in relation to their circumstances*) would have instrumental reasons to pursue the convergent instrumental goals. Because the conclusion of the argument isn't that ASIs would almost certainly try to act in ways that threaten the survival of humanity but merely that the possibility that they would try to do so is sufficiently likely or plausible to warrant our concern, even this more modest claim will suit its purposes.

2. Sharadin's Objection

In a recent paper in this journal, Nathaniel Sharadin argues that we "should not be worried"[10] by the instrumental convergence argument because we have insufficient reason to accept its crucial claim about the instrumental reasons of ASIs:

[9] Bostrom (2014), pp. 127, 131-32.
[10] Sharadin (2025), p. 1730.

Dangerous Convergent Instrumental Reason	For a wide range of possible ASIs with a wide range of final goals, those ASIs would have instrumental reasons to act so as to ensure their continued existence, retain their final goals, amplify their intelligence, improve their technology, and acquire resources without limit.[11]

Although this claim appears intuitively plausible at first glance, Sharadin argues that results from the philosophical literature on what it takes for something to promote the attainment of a goal show that we have insufficient reason to accept it. Thus, even if the rest of the argument is faultless, we needn't follow the argument to its conclusion because we needn't assent to this crucial step in it.

Sharadin thinks that the philosophical literature on promotion can shed light on whether we have sufficient reason to accept Dangerous Convergent Instrumental Reason because he thinks that facts about instrumental reasons are grounded in facts about promotion in the following way:

Promotionalism about Instrumental Reasons	An agent has an instrumental reason to perform a particular action if and only if (and because) its performing that action would promote the attainment of one of its goals.

Given that widely held principle about instrumental reasons, whose sufficient-condition (i.e., right-to-left) component I endorsed earlier, Dangerous Convergent Instrumental Reason is true if and only if the following claim is true:

Dangerous Convergent Promotion	For a wide range of possible ASIs with a wide range of final goals, the attainment of some of those goals would be promoted by the ASIs' acting so as to ensure their continued existence, retain their final goals, amplify their intelligence, improve their technology, and acquire resources without limit.

[11] This formulation differs from Sharadin's in two ways: (i) in accordance with what I wrote at the end of the previous section, it concerns only a wide range of ASIs with a wide range of final goals, not almost all possible final goals, and (ii) it doesn't focus exclusively on the act of acquiring resources without limit, as he does for the sake of simplicity. As I will explain later, insofar as his aim is to object to the argument as it is most charitably interpreted rather than as it is most ambitiously construed, this difference in formulations cast no doubt on my responses to his objection. Similar remarks will apply to my formulation of Dangerous Convergent Promotion below.

Although he concedes that Dangerous Convergent Promotion is intuitively plausible for reasons like the ones I gave earlier, he claims that "it is very important not to leave things at an intuitive level…. 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."[12] And if we survey philosophers' attempts to give a principled account of what it takes for something to promote the attainment of a goal, he argues, we will learn that we have insufficient reason to accept Dangerous Convergent Promotion.

As Sharadin notes, there are two kinds of philosophical theories of what it takes for the performance of an action to promote the attainment of a goal. On *probabilistic* theories, to promote the attainment of a goal is to increase, relative to some suitable baseline, the probability that it will be attained. On *fit-based* theories, to promote the attainment of a goal is to increase, relative to some suitable baseline, the degree to which the world resembles one in which it has been attained. Sharadin argues that, no matter which of these two types of theory we accept, and even if we accept some hybrid of the two, we will have insufficient reason to accept Dangerous Convergent Promotion. As he notes, however, Bostrom assumes a probabilistic theory of promotion: he writes, as we saw earlier, that "[s]everal instrumental [goals]… 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," and that "in many scenarios there will be future actions [an agent] could perform to *increase the probability* of achieving its goals."[13] For this reason, and because I agree with Sharadin that we have insufficient reason to accept Dangerous Convergent Promotion if we accept a fit-based account of promotion (and also because I find such accounts less independently plausible), I will focus on his argument concerning probabilistic theories of promotion.

The general idea that to promote the attainment of a goal is to increase the probability that it will be attained is plausible. It can explain, for example, why buying more lottery tickets promotes the goal of winning the lottery (which explains why someone who has the goal of winning the lottery has instrumental reason to buy more lottery tickets), and why wearing a seat belt while driving promotes the goal of avoiding injury (which explains why someone who has the goal of avoiding injury has instrumental reason to wear a seat belt while driving). However, philosophers have been unable to

[12] Sharadin (2025), p. 1733.
[13] Bostrom (2014), p. 132. Italics added.

produce a sufficiently precise formulation of it that isn't open to counterexamples. Because the probability of an event can qualify as *increased* only in relation to some baseline for comparison, a precise statement of this idea must specify what the relevant baseline is. But, as Sharadin observes, every baseline that has so far been proposed generates a probabilistic theory of promotion that fails to recognize certain cases of promotion and thus fails to recognize certain instrumental reasons.

Three baselines, and three corresponding probabilistic theories of promotion, have been proposed. The first baseline is how likely the attainment of the relevant goal would be if the agent were to *do nothing*, which generates the view that an agent's performing an action A promotes the attainment of a goal G if and only if it makes the attainment of G more probable than it would be if the agent were to do nothing.[14] This proposal initially seems unlikely to undercount cases of promotion, since it doesn't take much to make the attainment of a goal more probable than it would be if one were to do nothing: even buying a single ticket makes one more likely to win the lottery than one would be if one were to do nothing, for example, so this view rightly implies that buying a ticket promotes the goal of winning the lottery. However, this view implies that one cannot promote the attainment of a goal by doing nothing: after all, the probability that a given goal will be attained if one does nothing is just the probability that it will be attained if one does nothing, so doing nothing cannot make that probability any higher than it would be if one were to do nothing. But intuitively, doing nothing can promote the attainment of a goal: if a burglar who is motionlessly and silently hidden has the goal of avoiding detection, for example, then it appears that he could promote that goal by doing nothing. Relatedly, it seems that agents can sometimes have instrumental reason to do nothing.[15]

The second baseline is how likely the attainment of the relevant goal would be if the agent were *not to perform the relevant action*, which generates the view that an agent's performing an action A promotes the attainment of a goal G if and only if it makes the attainment of G more likely than it would be if the agent were not to do A.[16] The problem with this proposal is that it undercounts promotion (and thus instrumental reasons) in cases of overdetermination. Suppose that the plane and the train to New York are equally reliable, that my goal is only to get to New York (as opposed to getting there by a particular time), and that I would take the train if I were not to take the plane. Intuitively, taking the plane and taking the train would both promote my goal. However, the present proposal implies

[14] Schroeder (2007), p. 113.
[15] Evers (2009); Behrends and DiPaolo (2011).
[16] Finlay (2006), p. 8n16.

that taking the plane wouldn't promote my goal: after all, I would take the train if I were not to take the plane, and I would be no less likely to achieve my goal if I were to take the train than I would be if I were to take the plane. Although I appear to have instrumental reason to take the plane (though no more instrumental reason to do so than I have to take the train), this view implies otherwise.[17]

The third baseline is how probable the attainment of the relevant goal was *before* the agent performed the relevant action, which generates the view that an agent's performing an action A promotes the attainment of a goal G if and only if it makes the attainment of G more likely than it was before the agent did A.[18] This proposal has the problematic implication that, in certain kinds of situations, what seem to be merely less effective ways to promote the attainment of a goal are no ways to promote it at all. Suppose that there is a 10% chance that I will buy one lottery ticket and a 90% chance that I will buy ten tickets. It appears that each of the two actions would promote the goal of winning the lottery, though the former would promote it less well. The present proposal implies, however, that I would not promote my goal of winning the lottery if I were to buy one ticket. After all, if I were to do that, I would make it *less* likely that I will win the lottery than it was before I did that, when I still had a 90% chance of buying ten tickets instead of one. Although I seem to have instrumental reason to buy one ticket (though less instrumental reason to do so than I have to buy ten tickets), this view implies otherwise.[19]

These problems for these baselines are well known, and they have motivated some philosophers who understand promotion probabilistically to think of promotion as essentially *contrastive:*

Contrastive Probabilism about Promotion	The claim that an option O (e.g., an agent's performing action A) promotes the attainment of a goal G is always shorthand for the claim that O *rather than some other option O** promotes the attainment of G—where one option rather than another promotes the attainment of a goal if and only if the goal is *more likely* to be attained if the first option is realized than if the second one is.[20]

[17] Behrends and DiPaolo (2011).
[18] Lin (2018).
[19] Lin (2018), pp. 377-78.
[20] Sharadin (2025), p. 1738.

This view allows us to say the right things about the sorts of cases that bedeviled the three non-contrastive views. The burglar's doing nothing rather than waving his arms around promotes his avoiding detection; my taking the plane rather than the train doesn't promote my getting to New York, but my taking the plane rather than staying home does; and although my buying one ticket rather than buying ten doesn't promote my winning the lottery, my buying one ticket rather than buying none does. For these reasons, Sharadin claims that "if we're going to be *probabilists* about promotion, we should be *contrastive* probabilists."[21]

This is a problem for the instrumental convergence argument, Sharadin argues, because Dangerous Convergent Promotion assumes that promotion can be non-contrastive—that the attainment of a goal can be promoted by a particular option, and not just by a given option rather than some other option. Likewise, Dangerous Convergent Instrumental Reason assumes that instrumental reasons can be non-contrastive—that an agent can have instrumental reason to perform a particular action, and not just instrumental reason to perform a particular action as contrasted with some alternative to performing it. But Contrastive Probabilism about Promotion claims that promotion is always contrastive. Given Promotionalism about Instrumental Reasons, it also implies that instrumental reasons are always contrastive: an agent can only have instrumental reason to perform a particular action *A as contrasted with some alternative to performing it* (e.g., not performing *A*, or performing some other action *A**). Thus, Sharadin argues, the literature on promotion teaches us that the instrumental convergence argument presupposes an incorrect picture of promotion and instrumental reasons.

Of course, Sharadin concedes that one could recast the argument in a way that is compatible with contrastivism. This would involve restating Dangerous Convergent Instrumental Reason as follows:

Contrastive Dangerous Convergent Instrumental Reason (Contrastive DCIR)	For a wide range of possible ASIs with a wide range of final goals, those ASIs would have instrumental reasons to act so as to ensure their continued existence, retain their final goals, amplify their intelligence, improve their technology, and acquire resources without limit *as contrasted with alternative A to so acting*.

[21] Sharadin (2025), p. 1740.

But this is a schema that expresses different claims depending on how A is specified. And it's not clear, Sharadin argues, that there is any specification of A on which the truth of the resulting claim would suggest that ASIs would be likely to pursue the convergent instrumental goals—even if the rest of the argument goes through. Suppose, for example, that Contrastive DCIR is true when A is the ASIs' shutting themselves down. In that case, its truth would imply that a wide range of possible ASIs with a wide range of final goals would have instrumental reasons to pursue the convergent instrumental goals as contrasted with shutting themselves down. But because it's unlikely that ASIs would be forced to choose between acting in the former way and acting in the latter, this wouldn't suggest that they are likely to pursue the convergent instrumental goals—even if we can infer that ASIs are likely to do something from the fact that they have instrumental reasons to do it. At most, it would suggest that they are likely to pursue the convergent instrumental goals in the unlikely event that their only other option is to shut themselves down. In all likelihood, there are many other ways that ASIs could act, so the fact that they have instrumental reason to act in the former way rather than in the latter way doesn't suggest that they are likely to act in the former way. (It's plausible that you have instrumental reason to cut off one of your fingers rather than to cut off one of your hands. But since you needn't perform either act, this doesn't suggest that you will likely perform the first.)[22]

To summarize: Sharadin's objection is that the instrumental convergence argument depends on a view of promotion and instrumental reasons that we should reject. The philosophical literature on promotion teaches us that promotion is essentially contrastive and that instrumental reasons are, too. This means that Dangerous Convergent Promotion and Dangerous Convergent Instrumental Reason are false as we originally stated them, since they assume that promotion and instrumental reasons can be non-contrastive. And although one could reformulate the instrumental convergence argument in contrastivist terms, it is unclear how such a version of the argument could show that ASIs are likely to pursue the convergent instrumental goals—as opposed to merely being likely to pursue them in the unlikely event that they are forced to choose between pursuing them and acting in certain very specific other ways. Because the conclusion that ASIs are sufficiently likely to try to act in ways that are dangerous to humanity is supposed to follow from the claim that they are likely enough to pursue those instrumental goals, Sharadin argues, it is unclear how a contrastivist version of the argument could establish that conclusion.

[22] Sharadin (2025), pp. 1740-46.

3. Replies to Sharadin

I have two replies to Sharadin. The first is that, since he hasn't shown that promotion (if understood probabilistically) is essentially contrastive, he hasn't undermined the good reasons we have to accept Dangerous Convergent Promotion and Dangerous Convergent Instrumental Reason. The second is that, even if promotion is essentially contrastive, a contrastivist reformulation of the argument could plausibly show that, because of the instrumental reasons that they are likely to have, ASIs are likely to pursue the convergent instrumental goals.[23]

To understand my first reply, recall that Sharadin admits that Dangerous Convergent Promotion and Dangerous Convergent Instrumental Reason are intuitively plausible. With respect to acting so as to acquire resources without limit, for example, he writes:

> [A]t least intuitively, there doesn't appear to be any limit to the usefulness of acting to acquire more (and more, and more) physical resources…, no matter what one's goals, at least assuming that excess quantities of those resources will be exchangeable for some other useful goods…. 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.[24]

And although he doesn't explicitly discuss them, he would presumably say similar things about the other convergent instrumental goals. He would presumably agree, for example, that for a wide range of possible ASIs with a wide range of final goals, it is intuitively plausible that those ASIs' acting to ensure their continued existence would promote the attainment of those goals. After all, for a wide range of final goals, it seems that the continued existence of highly intelligent agents who have those goals would promote the attainment of those goals. Sharadin's claim isn't that we have no reasons to accept Dangerous Convergent Promotion or Dangerous Convergent Instrumental Reason; it is that

[23] A third possible response would be to reject the necessary-condition (i.e., left-to-right) component of Promotionalism about Instrumental Reasons: that an agent has an instrumental reason to perform a particular action *only if* its performing it would promote the attainment of one of its goals. If that claim is incorrect, then Dangerous Convergent Instrumental Reason could be true even if Dangerous Convergent Promotion is false, so the former claim could survive the success of Sharadin's contrastivist argument against the latter. But since it's unclear why we should accept Dangerous Convergent Instrumental Reason unless we accept Dangerous Convergent Promotion, I won't pursue this response.
[24] Sharadin (2025), p. 1733.

our intuitive reasons for accepting them, such as the ones just given, are undermined by results from the philosophical literature on promotion.

We can see that they are not so undermined, however, by getting clear on what that literature has and hasn't established. It is true, as Sharadin claims, that each of the three main attempts to make precise, in non-contrastivist terms, the idea that to promote the attainment of a goal is to increase the probability that it will be attained is open to counterexamples. It is also true, as he claims, that contrastive probabilism about promotion avoids those counterexamples. But this hardly establishes that "if we're going to be *probabilists* about promotion, we should be *contrastive* probabilists."[25] The fact that three attempts to precisely state an intuitively plausible idea are open to counterexamples doesn't establish, or even strongly suggest, that that idea isn't correct. It is intuitively plausible that promotion is not necessarily contrastive and that to promote the attainment of a goal is to increase the probability that it will be attained. If promotion is so understood, then Dangerous Convergent Promotion and Dangerous Convergent Instrumental Reason are intuitively plausible for reasons like the ones that I gave above and that Sharadin himself gives. The force of this intuitive case for those claims is not greatly diminished by the fact that the handful of philosophers who have so far tried to do so haven't yet succeeded at formulating non-contrastive probabilism in precisely the right way.

Furthermore, there seems to be something incoherent about Sharadin's argumentative strategy. He admits that Dangerous Convergent Promotion and Dangerous Convergent Instrumental Reason are intuitively plausible, but as I explained earlier, he says that "it is very important not to leave things at an intuitive level." Whether the instrumental convergence argument succeeds is a matter of practical importance: if it does succeed, for example, then governments might want to slow down or restrict research in AI. "These are not courses of action or policies that we should undertake on the basis of intuitive judgments about what promotes what."[26] But having thus claimed that intuitive judgments about promotion or instrumental reasons are insufficiently reliable, he then argues for contrastivism about promotion using judgments of precisely that kind. The problem that he highlights for each of the three existing non-contrastive forms of probabilism is that there are certain intuitive judgments that it cannot accommodate (e.g., that one can, in some situations, promote the attainment of a goal by doing nothing), and his case for contrastive probabilism is just that it can accommodate those

[25] Sharadin (2025), p. 1740.
[26] Sharadin (2025), p. 1733.

judgments. If, as he thinks and as I agree, our intuitions about promotion and instrumental reasons are sufficiently reliable to be the basis of strong arguments against the three existing non-contrastive forms of probabilism, then why aren't they also reliable enough to justify us in accepting, or at least in assigning fairly high credences to, Dangerous Convergent Promotion and Dangerous Convergent Instrumental Reason? He makes it seem as if he is evaluating those claims using standards that are better than those provided by intuitive judgments, but his case against them employs precisely the latter standards. Besides, as I just explained, it doesn't succeed even if we accept those standards because it only establishes the inadequacy of extant formulations of non-contrastive probabilism, not that of non-contrastive probabilism itself.

Moreover, remember that the three formulations of non-contrastive probabilism fail by *undercounting* cases of promotion and thus undercounting instrumental reasons: the problem is that they imply that there is no promotion where there intuitively is, and thus that there are no instrumental reasons where there intuitively are. If those views had failed in the opposite direction, by implying that there is promotion where there intuitively isn't (and thus that there are instrumental reasons where there intuitively aren't), then their failure might have given us reason to suspect that the correct account of promotion is a restrictive one on which promotion is relatively rare—an account whose truth might undermine Dangerous Convergent Promotion and Dangerous Convergent Instrumental Reason. Because those views fail by undercounting, however, their failure gives us no reason to suspect this. If anything, it highlights just how prevalent promotion and instrumental reasons are. Against this background, it seems perfectly credible that those claims are true if, as even Sharadin would agree, they are intuitively plausible.

To summarize: Sharadin has not succeeded in undermining the intuitive reasons that he admits we have to accept Dangerous Convergent Promotion and Dangerous Convergent Instrumental Reason. Although the truth of contrastive probabilism would indeed imply their falsity, he hasn't shown that contrastive probabilism is, or is even likely to be, true. This is because the familiar problems that he raises for the three existing non-contrastive forms of probabilism don't establish, or even strongly suggest, that non-contrastive probabilism is false: it would be premature to conclude that no non-contrastive form of probabilism can avoid those problems just because the three existing forms of it succumb to them. Moreover, because his argument against those three views assumes the reliability of the very sorts of intuitive judgments that suggest that Dangerous Convergent Promotion and

Dangerous Convergent Instrumental Reason are true, and because it highlights just how widespread promotion and instrumental reasons are, it inadvertently bolsters those claims.

Suppose, however, that I am mistaken about all this and that contrastive probabilism really is the correct account of promotion. In that case, I have a second response to Sharadin: contrary to what he argues, a contrastivist reformulation of the instrumental convergence argument could plausibly show that, because ASIs are likely to have certain instrumental reasons, they are likely to pursue the convergent instrumental goals.

Recall that, on such a reformulation, the argument would depend on Contrastive DCIR, which says that a wide range of possible ASIs with a wide range of final goals would have instrumental reasons to pursue the convergent instrumental goals *as contrasted with alternative A to pursuing them*. As I noted earlier, for *some* specifications of A (e.g., shutting themselves down), the truth of Contrastive DCIR wouldn't be worrisome, since it's unlikely that ASIs would need to choose between A, so specified, and pursuing those goals. What the argument needs is a specification of A on which Contrastive DCIR is true and, since what it says about the instrumental reasons of ASIs suggests (at least if the rest of the argument goes through) that they would likely pursue those goals, its truth is worrisome. Sharadin claims that "we don't have any reason to expect" that any such specification exists.[27]

It seems to me, however, that such a specification does exist. An obvious alternative to pursuing the convergent instrumental goals is *not pursuing them*. It appears that a wide range of possible ASIs with a wide range of final goals would have instrumental reason to pursue the convergent instrumental goals *as contrasted with not pursuing them*. After all, for a wide range of possible ASIs with a wide range of final goals (such as that of solving a particular set of unsolved problems in physics), it seems that those goals are more likely to be attained if those ASIs continue to exist, retain their final goals, are more intelligent, have better technology, and have ever larger quantities of resources than if they do not. And while some possible ASIs might, because of their limitations and circumstances, be unable to promote these convergent instrumental goals, a wide range of possible ASIs—ones that can make the attainment of these goals more likely if they pursue them than it would be if they didn't pursue them—could do so. If contrastivism is true, this implies that a wide range of ASIs with a wide range of final goals would have instrumental reason to pursue the convergent instrumental goals rather

[27] Sharadin (2025), p. 1742.

than not pursue them. But this fact is worrisome, at least if the rest of the argument goes through, since not pursuing those goals is just the negation of pursuing them. While it's unlikely that an ASI would be forced to choose between pursuing the convergent instrumental goals and shutting itself down, it is logically guaranteed that all ASIs will be forced to choose between pursuing those goals and not pursuing them. If, as I have argued, Contrastive DCIR is true when A is so specified, then provided that the rest of the argument goes through, its truth would indeed support the claim that that ASIs would likely pursue the convergent instrumental goals.

We should dwell for a moment on how capable an ASI would have to be, relative to the situation in which it finds itself, to be among the ASIs that have instrumental reasons to pursue the convergent instrumental goals as contrasted with not doing so. On one end of the spectrum, we can imagine an ASI whose thoughts are completely transparent to a team of overseers who would successfully shut it down as soon as it forms an intention to pursue any of those goals. Presumably, even if this ASI's final goals are more likely to be attained if those convergent instrumental goals are *attained* than if not, it would have no instrumental reason to *pursue* the latter goals as contrasted with not pursuing them, since none of the actions available to it would be such that its performing that action rather than not performing it would promote the latter goals. On the other end of the spectrum, we can imagine a god-like ASI whose thoughts are inscrutable and whose "superpowers"[28] make it highly probable that it will attain any goal it pursues. If this ASI's final goals are more likely to be attained if the convergent instrumental goals are attained than if not, then since the latter goals are more likely to be attained if it pursues them than if not, it will have instrumental reasons to pursue the latter goals rather than not pursuing them. Now, if it were true that only such god-like ASIs would have such instrumental reasons, that might be reassuring, since it might seem unlikely that ASIs will be god-like in those ways.[29] But that doesn't seem true. Instead, it seems that a broad range of ASIs in the middle of the spectrum of capabilities could promote those instrumental goals by pursuing them rather than not pursuing them and would thus have instrumental reasons to pursue them rather than not pursuing them. One needn't be god-like to promote one's continued survival by acting so as to ensure it as contrasted with not so acting: it need only be the case that it is *more likely* that one will survive if one so acts than if not. There are many possible ASIs with varying levels of less-than-god-like power for which that comparative probability claim is true—e.g., ones with significant though

[28] Bostrom (2014), pp. 113-14.
[29] Though see Bostrom (2014), ch. 6 for an argument that this isn't as unlikely as it might seem.

imperfect abilities to deceive us regarding their plans, in situations with only moderately competent human oversight. The same seems true of the rest of the convergent instrumental goals. Thus, it is plausible that a wide range of possible ASIs with a wide range of final goals *and a wide range of abilities*, and not just the narrow subset of these with god-like abilities, would have instrumental reasons to pursue the convergent instrumental goals as contrasted with not pursuing them. Since this would indeed suggest that such a wide range of ASIs would likely pursue those goals—at least if the rest of the argument, which Sharadin doesn't criticize, goes through—the truth of contrastivism wouldn't undermine the instrumental convergence argument.

Some of Sharadin's remarks indicate that he would deny that a wide range of possible ASIs with a wide range final goals would have instrumental reason to pursue the convergent instrumental goals instead of not pursuing them. Focusing on the pursuit of resources without limit, he writes:

> 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. Instead, many goals are equally likely to be achieved given moderate levels of resources as they are given extreme ones. 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.[30]

Although the alternative to the unlimited pursuit of resources that he considers here is the moderate (and thus limited) pursuit of resources, he would presumably make the same argument regarding the alternative of not unlimitedly pursuing resources. There are many goals, he would maintain, that give those who have them no instrumental reason to pursue resources without limit as contrasted with not doing so, because there are many goals (e.g., buying a stick of gum at the corner store) such that, if you already have certain moderate amounts of resources, their attainment is no more likely if you acquire more resources than if you don't.

[30] Sharadin (2025), p. 1745.

I think he is mistaken about this, however. Even if we assume that a stick of gum costs less than a dollar at the corner store, you are at least slightly more likely to attain the goal of buying a stick of gum from the store if you have a million dollars than if you have only one. After all, there is some chance that you will be robbed of your dollar or lose it on your way to the store. This would prevent you from attaining your goal if you have only one dollar, but not if you have a million. As Bostrom notes, even if a particular goal requires only a modest amount of resources, one can usually increase the probability that one will attain it by having resources beyond that amount, and it is plausible that there is no amount of resources, however high, at which this ceases to be true.[31]

Although Sharadin doesn't do this in the passage above, he elsewhere distinguishes, as I did earlier, between (i) whether a final goal is more likely to be attained if a given instrumental goal is *attained* than if it isn't, and (ii) whether a final goal is more likely to be attained if one *pursues* or *tries to attain* a given instrumental goal than if one doesn't.[32] Even though your goal of buying the stick of gum is more likely to be attained if you *have* a million dollars than if you don't, it's plausibly no more likely to be attained if you *try to acquire* a million dollars than if you don't. For this reason, it's plausible that it wouldn't be promoted by your trying to acquire a million dollars rather than not trying to do this. However, this doesn't establish that a goal would be promoted by one's trying to acquire resources without limit (as contrasted with not trying to do this) only if the goal *requires* more resources than one has, as the passage from Sharadin suggests: since the likelihood that a goal will be attained can increase with the amount of resources that one has even if that goal doesn't require more resources than one has, agents that can effectively pursue resources that they don't already have can thereby promote the attainment of even such a goal. Nor does it undermine the reasons I gave earlier for thinking that a wide range of possible ASIs with a wide range of final goals could promote their final goals by trying to acquire resources without limit rather than not so trying: an ASI whose final goal is to solve certain problems in physics could, for example, promote the attainment of this goal by trying to acquire resources without limit (as contrasted with not so trying) if, because of how capable it is in relation to its circumstances, it is more likely to obtain more resources than it already has if it tries to do this than if it doesn't. Thus, although there are certain goals (e.g., that of buying a stick of gum) whose attainment certain agents (e.g., people who can already afford a stick of gum) plausibly wouldn't promote by trying to acquire more resources rather than not so trying, it remains plausible

[31] Bostrom (2014), pp. 137-38, 149-52.
[32] Sharadin (2025), pp. 1752-53. Also see Gallow (2025), p. 1585.

that a wide range of possible ASIs with a wide range of final goals would have instrumental reasons to try to acquire resources without limit as contrasted with not so trying. The same is true of the rest of the convergent instrumental goals. But if this is so, then since the claim that a wide range of ASIs with a wide range of final goals would have instrumental reasons to pursue those goals as contrasted with not pursuing them would indeed suggest that they are likely to pursue them—at least if the rest of the argument, to which contrastivism provides no objection, goes through—contrastivism does not undermine the instrumental convergence argument.

Sharadin might respond that, because "Bostrom frames his argument as one that remains agnostic concerning the specific content of [possible ASIs' final] goals,"[33] and because I haven't explained how a contrastivist version of the argument could show that ASIs would have instrumental reasons to pursue the convergent instrumental goals "[a]lmost whatever"[34] their final goals are, I haven't answered his objection. As I emphasized earlier, however, although Bostrom does at times suggest that we can infer that ASIs would have instrumental reasons to pursue the convergent instrumental goals even if we are *almost completely* agnostic about the content of their final goals, his argument is more charitably construed as claiming only that we can infer this even if we are *moderately* agnostic about this—i.e., even if we suppose only that their final goals would fall within a particular "wide range"[35] in the space of possible goals. After all, even this less ambitious version of the argument would, if successful, show that the probability that ASIs would try to act in ways that threaten the survival of humanity is sufficiently high to warrant our concern.[36] Thus, even if Sharadin has shown that the most ambitious form of the argument fails, he hasn't refuted the version of the argument that is most plausible and most worthy of our attention.

---

[33] Sharadin (2025), p. 1746.
[34] Sharadin (2025), p. 1740.
[35] Bostrom (2014), pp. 127, 132.
[36] Some might doubt this claim. If, as the version of the argument that I have focused on allows, there are many possible final goals that *wouldn't* give ASIs that have them instrumental reasons to pursue the convergent instrumental goals, then couldn't we just build only ASIs whose final goals wouldn't give them such instrumental reasons? Doesn't this show that it would be rather *unlikely* that ASIs would try to do things that threaten the survival of humanity, because we would be unlikely to give them the sorts of final goals that the argument warns us it would be dangerous for them to have? This train of thought assumes something dubious, however: that we would have enough control over the final goals of ASIs to make it highly probable that they have only the final goals that we intend them to have. It also assumes, perhaps even more dubiously, that it's unlikely that anyone in a position to build an ASI would intend for it to have any final goals of the sort that the argument deems dangerous. Even if the argument were disseminated far more widely than it is realistic to expect, there are many reasons (e.g., financial incentives, ease of programming) why some people would aim to build an ASI with final goals of the sort that, according to the argument, generate dangerous convergent instrumental reasons. See Bostrom (2014), pp. 129, 143-44.

In summary, my second reply to Sharadin is that a contrastive form of the instrumental convergence argument could plausibly show something about the instrumental reasons of ASIs that suggests—at least if the rest of the argument (which he doesn't criticize) goes through—that they are sufficiently likely to pursue the convergent instrumental goals and to thereby threaten the survival of humanity. Thus, even if he is right that contrastive probabilism is the correct account of promotion and thus that instrumental reasons are essentially contrastive, this doesn't undermine the argument.

4. Motivations, Capabilities, and Instrumental Convergence

Let me end by emphasizing a general point that has implicitly come up earlier.

It might seem that we can cleanly separate and independently address questions about the *motivations* of ASIs (e.g., whether they would aim or attempt to eliminate or otherwise greatly harm humanity), on the one hand, and questions about their *capabilities* (e.g., whether they would be able to do that if they were to aim or attempt to do it), on the other, and that the instrumental convergence argument concerns only questions of the former sort. This impression seems corroborated by the fact that, as I noted at the outset, the argument purports to show only that ASIs would be sufficiently likely to attempt to do certain dangerous things, not that they would likely succeed in doing them. It might also be encouraged by Bostrom's decision to present the argument in a chapter on the motivations of ASIs instead of in an adjacent chapter on their capabilities, and it might seem to derive further confirmation from Bostrom's claim that there are two classes of methods for trying to prevent an ASI from causing an existential catastrophe: *capability control* methods, which "seek to… limit[] what the superintelligence can do," and *motivation selection* methods, which "seek to… shap[e] what the superintelligence wants to do."[37]

On reflection, however, the truth is more complicated. Although the conclusion of the instrumental convergence argument concerns only what ASIs would likely aim or attempt to do, the argument does make some assumptions about what capabilities ASIs would likely have. This is because what actions an agent has instrumental reasons to perform depends on its capabilities and circumstances, which together determine which, if any, actions available to it would promote the attainment of its final goals. The less powerful an agent is in relation to its circumstances, the fewer opportunities it

[37] Bostrom (2014), pp. 157, 169.

has to promote the attainment of its final goals, and the fewer actions it has instrumental reasons to perform. Thus, the argument isn't compatible with every conceivable view about the capabilities of ASIs: assuming that all ASIs would be at the near end of the spectrum of capabilities that I sketched earlier, it couldn't show that they would likely pursue the convergent instrumental goals and, in so doing, threaten the survival of humanity. I argued earlier that the argument needn't make implausibly strong assumptions about the capabilities of ASIs (e.g., that they would all be god-like in power): it need only deem it sufficiently likely that they would be capable of increasing the probability that the convergent instrumental goals will be attained, which they could do even without god-like powers if our oversight of them is realistically imperfect. As modest as it is, however, this is an assumption about what capabilities ASIs would likely have, and the argument requires some such assumption. One promising avenue for future research would be to precisely state and assess what the argument must assume about the likely capabilities of ASIs.

These considerations show that, if the instrumental convergence argument is on the right track, then advanced capabilities in ASIs would be doubly dangerous. They would be dangerous because they would increase the likelihood that ASIs would succeed in eliminating humanity if they were to aim or attempt to do so. Less obviously, however, they would also be dangerous because, by increasing the likelihood that ASIs would have instrumental reasons to eliminate humanity, they would increase the likelihood that ASIs would aim or attempt to do this. Thus, if we could limit the capabilities of ASIs (e.g., by preventing them from coming anywhere close to having "superpowers" in the realms of intelligence amplification, strategizing, social manipulation, hacking, technology research, and economic productivity[38]), this would reduce both their ability to threaten the survival of humanity and their motivation to do so. For this reason, capability control methods are even more important than they might have initially seemed.

5. Conclusion

The instrumental convergence argument says that, since a wide range of possible ASIs with a wide range of final goals would have instrumental reasons to pursue certain convergent instrumental goals the pursuit of which would threaten the survival of humanity, the probability that ASIs would try to act in ways that greatly endanger humanity is high enough to merit our concern. I have defended this

[38] Bostrom (2014), p. 114.

argument against Sharadin's objection that the philosophical literature on promotion shows that we have insufficient reason to accept its claims about ASIs' instrumental reasons. Not only has he not shown the falsity of the picture of promotion and instrumental reasons that the argument assumes, but even if he had, a suitably reformulated version of the argument would withstand his criticisms. Obviously, however, this is not a vindication of the argument as a whole. Indeed, even the part of the argument that he targets—its claim about the instrumental reasons of a wide range of ASIs with a wide range of final goals—may be problematic for other reasons.[39] Whether the argument is sound is a matter that merits further investigation.

I have also made two more general points about the argument. First, rather than trying to show that ASIs would (or would almost certainly) try to do things that threaten the survival of humanity, the argument tries to show only that the possibility that they would try to do such things is sufficiently likely or plausible to merit our concern. Second, although the argument doesn't purport to *establish* anything about what ASIs would be capable of doing (as contrasted with what they would aim or attempt to do), it does *assume* something about what ASIs would be capable of doing. Relatedly, I have argued that, if the argument succeeds, then capability control methods are doubly important for preventing ASIs from causing an existential catastrophe: if effective, they make ASIs not only less likely to be able to eliminate humanity but less likely to aim or attempt to do so.

[39] See Gallow (2025) and Southan et al. (forthcoming). These authors take the argument to be concerned with what it would be instrumentally *rational* for ASIs to do rather than what ASIs would have instrumental *reasons* to do, however. It's not immediately clear how well their objections would carry over to the reasons-based reading of the argument that Sharadin and I accept, which is at least as well supported by Bostrom's remarks as the rationality-based one. Moreover, Gallow's target is something like the version of the argument that I've deemed overly ambitious.

Works Cited

Behrends, Jeff and DiPaolo, Joshua (2011), "Finlay and Schroeder on Promoting a Desire," *Journal of Ethics & Social Philosophy*.

Bostrom, Nick (2014), *Superintelligence* (Oxford: Oxford University Press).

Evers, Daan (2009), "Humean Agent-Neutral Reasons?" *Philosophical Explorations* 12: 55-67.

Gallow, J. Dmitri (2025), "Instrumental Divergence," *Philosophical Studies* 182: 1581-1607.

Lin, Eden (2018), "Simple Probabilistic Promotion," *Philosophy and Phenomenological Research* 96(2): 360-79.

Omohundro, Stephen (2008), "The Basic AI Drives," in Pei Wang, Ben Goertzel, and Stan Franklin (eds.), *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (Amsterdam: IOS Press).

Schroeder, Mark (2007), *Slaves of the Passions* (Oxford: Oxford University Press).

Sharadin, Nathaniel (2025), "Promotionalism, Orthogonality, and Instrumental Convergence," *Philosophical Studies* 182: 1725-55.

Southan, Rhys, Ward, Helena, and Semler, Jen (forthcoming), "A Timing Problem for Instrumental Convergence," *Philosophical Studies*.