Are Current AI Systems Capable of Well-Being?

James Fanciullo

Lingnan University

Forthcoming in the *Asian Journal of Philosophy*

Abstract: Recently, Simon Goldstein and Cameron Domenico Kirk-Giannini have argued that certain existing AI systems are capable of well-being. They consider the three leading approaches to well-being—hedonism, desire satisfactionism, and the objective list approach—and argue that theories of these kinds plausibly imply that some current AI systems are capable of welfare. In this paper, I argue that the leading versions of each of these theories do not imply this. I conclude that we have strong reason to doubt that current AI systems are capable of well-being.

Are any current artificial intelligence (AI) systems capable of well-being or welfare? In their interesting paper "AI Wellbeing," Simon Goldstein and Cameron Domenico Kirk-Giannini answer Yes. Goldstein and Kirk-Giannini are led to this surprising answer by considering leading theories of mental states such as desire and belief, as well as leading theories of the nature of well-being, and suggesting that many of these theories together plausibly imply that AI systems of a certain kind are welfare subjects. Since the AI systems that Goldstein and Kirk-Giannini are focused on in fact already exist, their arguments imply that some existing AI systems have well-being—or, that things can be intrinsically good or bad *for* the systems, or that the systems' lives (or perhaps more accurately, existences) can go better or worse *for them*. Given these systems that are putatively capable of welfare also seem likely to become increasingly common in the near future, the authors' arguments are not just surprising, but also potentially quite pressing, if they're successful.

The kind of AI system Goldstein and Kirk-Giannini have in mind is what they call an *artificial language agent* (or *language agent*, for short). These agents have as their core a *large language model*, or LLM, which is "an artificial neural network designed to generate coherent text responses to text inputs" (p. 3). Artificial language agents "augment large language models (LLMs) with the capacity to observe, remember, and form plans" (p. 2).[1] Notably, then, these language agents represent a relatively incremental advancement on LLMs. They need not be, for example, phenomenally conscious, and Goldstein and Kirk-Giannini's arguments therefore do not require that the agents be phenomenally conscious for them to be capable of welfare.[2] These language agents need not be—and those that presently exist presumably are not—so sophisticated. Instead, as Goldstein and Kirk-Giannini put it: "Language agents are built by wrapping an LLM in a larger functional architecture that allows the system to engage in long term planning" (p. 3). To illustrate how these agents work, the authors offer a useful analogy with a human being and their cerebral cortex. A human being is not identical to their cerebral cortex, though the cerebral cortex performs most of the human's cognitive processing. Likewise, a language agent is not identical to its LLM, though its LLM performs most of its processing. Moreover, just as a human being stores and retrieves information regarding their beliefs and desires, appeals to this information in making plans, etc., so the language agent stores and retrieves information (in the form of natural-language sentences) regarding its "beliefs" and "desires," feeds this information to its LLM which produces a "plan" as an output, etc.

[1] All page numbers in the text will refer to Goldstein and Kirk-Giannini forthcoming, unless otherwise noted.

[2] Goldstein and Kirk-Giannini do not take a stand on whether language agents are phenomenally conscious. Here, however, I will assume, in accordance with what I take to be widespread intuition, that they are not. For arguments in favor of this, see Butlin et al. 2023. Though see also Goldstein and Kirk-Giannini 2025.

These observations suggest that the processes used by language agents to store and retrieve information, make plans, etc., "obey the familiar laws of folk psychology" (p. 4).

As I say, Goldstein and Kirk-Giannini go on to consider leading theories of belief and desire. They conclude that "a wide range of accounts of the nature of belief and desire entail that systems like language agents can have beliefs and desires" (p. 9). While I think their arguments here can also be questioned, my focus in this paper will instead primarily be on their discussion of the nature of well-being. Together with their conclusions regarding desire and belief, Goldstein and Kirk-Giannini take their discussion of the leading approaches to welfare—hedonism, desire satisfactionism, and the objective list approach—to show that language agents can have welfare as well. In what follows, I'll consider their arguments regarding these theories of well-being in turn, and in each case argue that we have some significant reasons to doubt their claims about language agents' capacity for welfare.

1. The three leading approaches to welfare

1.1 Hedonism

According to *hedonism*, all and only enjoyable (or pleasurable) experiences are basically good for one, and all and only unpleasant (or painful) experiences are basically bad for one. Since experiences of enjoyment and unpleasantness seem to require consciousness, hedonism seems to imply that things incapable of consciousness are also incapable of well-being. This, of course, includes language agents, which—at least among those that currently exist—certainly do not seem capable of consciousness.[3] Hedonism thus seems to imply that existing language agents do not have any capacity for welfare, which is of course at odds with Goldstein and Kirk-Giannini's surprising conclusion.

[3] For discussion of the possibility of suffering in digital systems, see Saad and Bradley 2022.

Admittedly, Goldstein and Kirk-Giannini readily admit this: they explicitly assume that if hedonism is true, language agents lack a capacity for welfare (p. 10). While they hint at a brand of hedonism that denies that experiences of enjoyment and unpleasantness require consciousness—one built on the idea that enjoyment and unpleasantness may be explained in terms of desire and belief, which themselves do not require consciousness (or so the view would presumably go)—they stop short of offering it in full. I'll similarly stop short of speculatively discussing their view, given they haven't offered it. Still, I will briefly say this. Any view that claimed a life could be good or bad for its subject without the subject experiencing any positively valenced, affectively or phenomenologically salient representations or experiences would not seem to me to live up to the name "hedonism." And so, any theory implying that existing language agents—which are presumably incapable of such experiences—are capable of welfare would not seem to live up to that name either. It therefore seems to me, to borrow one of the authors' own phrases, that Goldstein and Kirk-Giannini's admission that hedonism is inconsistent with their view that existing language agents are capable of welfare "is as it should be" (p. 8).

1.2 Desire satisfactionism

Desire satisfactionist theories claim that well-being consists, somehow or other, in the satisfaction of desire. And, at least recently, they've been taken to claim that ill-being consists, somehow or other, in the satisfaction of aversion (desire's negative counterpart).[4] There are many forms of desire satisfactionism, corresponding to the number of ways the "somehow or other" clause in the view can be spelled out. Now, as I've mentioned, Goldstein and Kirk-Giannini argue that several leading theories of desire and belief imply that existing language agents are capable of having these attitudes. Given this conclusion—which, again, I will simply grant here—one might expect that many leading

[4] See e.g. Fanciullo forthcoming and Pallies 2022.

versions of desire satisfactionism will straightforwardly imply that language agents are capable of welfare, too. Here, though, the details of the desire satisfactionist view seem to matter. Importantly, these days, few of those theorizing about the nature of well-being seem to accept the simple, "unrestricted" version on which the (objective, actual) satisfaction of just any of our desires makes us better off.[5] While this view would present perhaps the most straightforward route to Goldstein and Kirk-Giannini's desired conclusion, it has also allegedly been shown to have a variety of counterintuitive implications, leading to many attempted refinements and alternative versions of the approach.[6] Unfortunately, moreover, most all of these alternative versions have been shown to have various counterintuitive implications of their own.[7] This seemingly leaves us with little consensus as to the most plausible version of desire satisfactionism, making it difficult to assess whether any leading version of desire satisfactionism—outside of those, like "unrestricted" versions, with well-known problems—is consistent with Goldstein and Kirk-Giannini's claim that current language agents are capable of welfare.

Importantly, however, one recently proposed version of desire satisfactionism seems to have considerably advanced debate over the approach, if not established itself as the leading contender. This is the view offered by Chris Heathwood, who distinguishes between two kinds or senses of "desire."[8] Specifically: sometimes when we desire things, we're not just disposed to bring the things about, but are genuinely attracted to the things—we're engaged or compelled by them, or think of them with excitement or gusto. Take, for example, the desire for one's favorite meal when one is

[5] See e.g. Heathwood 2016 and Heathwood 2019. Of course, this is not to say that this version of the theory is unpopular among theorists of other kinds, such as economists and decision theorists (see e.g. Heathwood 2016 and Fanciullo 2022). However, the fact that theorists of well-being specifically seem to have largely abandoned it—or at least largely agree it stands in need of revision—seems to me strong evidence that the "unrestricted" version of desire satisfactionism should not be considered a leading theory of welfare.

[6] There are far more examples than I can reasonably list here, but for a start, see e.g. Heathwood 2016 and Heathwood 2019.

[7] Again, see e.g. Heathwood 2016 and Heathwood 2019.

[8] Heathwood 2019.

hungry, or the desire to see one's partner after a long time apart. In these cases, we're not just disposed to get the things, but we see something positive in them, regard them in a positive way, or see them under some favorable mode of presentation. Call desires of this sort *genuine* desires. In contrast, other times when we desire things, we're merely disposed to bring the things about. Consider, for example, going to the dentist despite loathing the thought of doing so, or being dragged to yet another pointless meeting. So long as one does these things, it seems one must have been disposed to do them, and must have in that sense desired to do them. Still, doing these things clearly does not require any attraction or engagement on the part of the person doing them. The person instead merely has a bare disposition to act. Call desires of this sort *behavioral* desires. According to Heathwood, it is only genuine desires, and not behavioral desires, that are relevant to well-being. Thus, whereas it may make us better off to get what we genuinely desire, it makes us no better off to get what we behaviorally desire.

As one may expect, this move allows the desire satisfactionist to avoid a number of well-known objections to alternative versions of the view. One such objection involves cases of compulsive desires. Consider, for instance, Warren Quinn's *Radio Man*, who is "in a strange functional state that disposes [him] to turn on radios that [he sees] to be turned off" (Quinn 1993, p. 32). As I've recently explained elsewhere:

> This state, which is said to be a desire, is meant to be *merely* functional, in that it is no more than a bare disposition to behave. There is no further purpose for which he turns on the radios, and indeed nothing favourable he sees in anything relating to turning them on (not hearing music, flipping switches, or anything else). He is in a motivational state that is, in effect, stipulated *not* to involve any attraction or engagement. (Fanciullo 2025, p. 50)

The worry for the standard desire satisfactionist, of course, is that when Radio Man turns on a radio, he does not thereby seem to be made better off. This is in spite of the fact that he performs a motivated act, and thereby satisfies one of his desires. It thus seems that well-being cannot consist in the satisfaction of just any of our desires. On Heathwood's account, in contrast, our judgment about Radio Man is entirely expected. Given Radio Man is stipulated not to be genuinely attracted to or engaged by turning on radios, and is instead merely disposed to act, his desire to turn on radios must be merely behavioral, in which case his desire here is not relevant to his welfare. Moreover, if we imagined, contrary to the case, that Radio Man *was* attracted to or engaged by the prospect of turning on radios, his desire to turn them on suddenly *would* seem relevant to his welfare (or so we might plausibly think). Heathwood's account thus seems to get the correct result in these, as well as other, important cases.[9] And, indeed, his account seems to illuminate something important about the connection between desire and welfare.

For Goldstein and Kirk-Giannini, then, it would be ideal if Heathwood's account, too, were to imply that current language agents are capable of welfare. And while this may at first glance seem unlikely, given the account's focus on the difference between mere behavior and genuine attraction, Goldstein and Kirk-Giannini argue that Heathwood's account in fact leaves room for current language agents who are capable of welfare. For this to be the case, of course, current language agents must be capable of genuine desire. And so, the question is whether language agents are so capable. To argue that they are, Goldstein and Kirk-Giannini offer what they take to be a way of distinguishing genuine desires from compulsive desires. In cases of compulsion, they suggest, an agent does not become disposed to perform an action by performing instrumental reasoning toward promoting their goals. Instead: "an agent's disposition to act is produced directly by some identifiable factor, such as a chemical addiction, in a way that is not appropriately sensitive to

[9] See Heathwood 2019.

processes like practical deliberation" (p. 11). In contrast, in cases of genuine attraction, they suggest, an agent becomes disposed to perform an act by performing instrumental reasoning toward achieving their goals. If that's right, then, given current language agents seem capable of becoming disposed to perform acts by performing such instrumental reasoning (or so we'll assume), it seems language agents may be capable of genuine attraction. And so, it seems language agents are capable of well-being, even on Heathwood's more sophisticated desire satisfactionist view.

Two things are worth noting here. First, whereas Goldstein and Kirk-Giannini focus on the distinction between genuine attraction and compulsion, there seems to be a third sort of phenomenon that's also important to distinguish here: non-compulsive, yet non-attracted, acts. Consider again, for example, going to the dentist despite loathing the thought of doing so, or being dragged to yet another pointless meeting. Surely these acts need not be compulsive—they may be produced in a way that's properly sensitive to processes like practical deliberation—but nor must they be cases of genuine attraction. Indeed, these are precisely our paradigm cases of action without genuine attraction. Given there are cases of this sort, then, it seems becoming disposed to perform an act by performing instrumental reasoning toward achieving goals is not sufficient for genuine attraction. And so, even if current language agents can become disposed to act by performing such reasoning, this does not imply that they are capable of genuine desire.

This brings us to the second thing worth noting. Given that becoming disposed to perform an act by performing appropriate instrumental reasoning is not sufficient for being genuinely attracted, it seems something *else* is needed—perhaps in addition to becoming disposed to act by performing such reasoning—to ensure genuine attraction. The questions then become: (i) what is this something else; and (ii) are current language agents capable of it?

Regarding (i), perhaps the most natural answer, and the one I find difficult to resist, is that genuine attraction requires some form of phenomenologically salient, positively valenced affective experience.[10] Indeed, what seems to make an attraction "genuine"—the reason we use this descriptor—is that there is something phenomenologically salient about it. It is this positive (psychological) feeling or phenomenological salience that makes it *true* or *genuine* attraction. Incidentally, the same seems true of aversion: we are *genuinely* averse to something when there's some negative "way it's like" to see the thing—when there's some negative mode of presentation under which we see it. If there were no such "way it's like," either positive or negative, there would seemingly be no reason to distinguish between being disposed to (not) do something and being *genuinely* attracted (or *genuinely* averse) to doing it. To see this, we can notice that this phenomenological salience seems to mark the difference between, for example, merely being disposed to go to the meeting, and being attracted to or enthused about going. Moreover, and importantly, this phenomenological salience seems to make sense of our intuitions about well-being in our main case of interest, namely Radio Man. The presence of this phenomenological salience seems to explain why, when we imagine Radio Man as being genuinely attracted to turning on radios, turning on radios plausibly makes him better off. And the lack of this phenomenological salience seems to explain why, when we imagine Radio Man with a bare disposition to act, turning on radios plausibly makes him no better off. The ultimate explanation here is not, then, merely that only in the former case does Radio Man become disposed to act via appropriate instrumental reasoning. Instead, it seems to be that only in the former case is turning on radios an act with phenomenologically salient, positively valenced affective appeal. Or so it seems to me. Frankly, I struggle to think of any other plausible candidate for the "something else" that's needed (in addition to becoming disposed to act through performing instrumental reasoning, assuming that's needed as well) to ensure *genuine* attraction. If my thought here is right, then, there is a positive "way it's like" to be genuinely attracted, or to genuinely desire. To genuinely desire is, in part, to (be disposed to)

[10] See Fanciullo 2025, Railton 2012, and Smithies forthcoming.

experience some phenomenologically salient, positive affect, or some positive affective representation of the desire's object.

Regarding (ii), it seems clear to me that current language agents are not capable of such phenomenologically salient experience. This experience seems to require phenomenal consciousness, which current language agents don't seem capable of. Thus, if my arguments here are correct, current language agents aren't capable of genuine desire, and so aren't capable of welfare on this version of desire satisfactionism.

On what I take to be the most plausible understanding of the most plausible version of desire satisfactionism, then, the view implies that current language agents are incapable of welfare. Admittedly, of course, you might reject either this version of the view or my understanding of it. And, in that case, I will have gone little way toward showing that your preferred version of desire satisfactionism is inconsistent with the conclusion that current language agents are capable of welfare. As I see it, however, insofar as we think desires are relevant to well-being, we have overwhelming reason to accept Heathwood's distinction: if we flatly ignore it, our version of desire satisfactionism will have enough problems already—language agents can simply be added to the list. And, insofar as we think it's possible that some acts are neither compulsive nor genuinely attractive, we have strong reason to think that the crucial missing feature in these acts is any phenomenologically salient, affective appeal on the part of those performing them. It seems, then, that as far as the connection between desire satisfaction and welfare goes, we have strong reason to accept two further claims: that phenomenal consciousness is required for welfare, and that current language agents are incapable of welfare.

1.3 Objective list theories

According to *objective list theories*, well-being consists in the possession of certain "objective goods." There are of course a variety of proposed objective lists, though goods commonly appearing on these lists include friendship, happiness, knowledge, achievement, rationality, and the like.[11] For Goldstein and Kirk-Giannini's purposes, then, the relevant question is whether current language agents are capable of possessing these goods. Focusing on two of these goods—rationality and knowledge—as test cases, the authors answer Yes.[12]

Regarding rationality and knowledge, it seems plausible enough that current language agents may know things, and moreover that they display a wide range of reasoning abilities. Again, we may doubt whether language agents are in fact capable of these things—for instance, whether they're truly capable of belief, or whether they really reason rather than just seeming to—but I'll set these questions aside.[13] My concern here is instead about whether these things should be regarded as basic constituents of our well-being in the first place. Notably, I'm not the first person to press the following concern, and even leading objective list theorists have admitted its force.[14] The problem is that any proposed objective good that does not properly connect with an agent's positive evaluations or pro-attitudes may for that reason seem incapable of being a basic good for the agent. Theories proposing objective goods of this sort are said to be "intolerably alienating," as they imply that a

[11] See Fletcher 2016.

[12] The authors also consider achievement, though their arguments in that case appeal to a more specific version of the objective list approach, namely "perfectionism." Unfortunately, I simply lack the space here to explain and respond to their arguments regarding perfectionism. However, I should briefly note, first, that perfectionism seems to have even more extreme implications regarding language agents than the authors' discussion may suggest—seeming to imply that even a language agent programmed to continuously add 1 to some sum may achieve an arbitrarily high level of welfare in the matter of minutes—and second, that we seem to have strong independent reasons for rejecting perfectionism as an approach to well-being. See e.g. Dorsey 2010 and Kitcher 1999. Relatedly, note that perfectionist views may be objected to on the grounds that they are "intolerably alienating" in the sense discussed below. To be clear, then, unlike in the cases of hedonism, desire satisfactionism, and the (non-perfectionist) objective list approach, I do not deny that (the best version of) perfectionism implies that existing language agents are capable of welfare. However, if proponents of the view that existing language agents are capable of welfare must pin their hopes on perfectionism—given my arguments here regarding the alternative approaches to welfare—then I believe opponents of this view can rest relatively easy.

[13] For discussion of whether LLMs are capable of belief, see Levinstein and Herrmann 2024.

[14] See e.g. Fletcher 2013 and Fletcher 2016.

subject's life may go well for them despite their having no positive evaluations or pro-attitudes towards anything in their life at all.[15] As a leading objective list theorist, Guy Fletcher, has noticed, any proposed objective good that does not itself require positive evaluations or pro-attitudes on the part of the subject who possesses it seems susceptible to this objection.[16] And so, it seems the most plausible versions of the objective list theory will be those including on their list only goods that require some form of pro-attitude, positive evaluation, or engagement.[17] Since a person could be highly skilled at reasoning or have massive amounts of knowledge—and so have a life that is very high in welfare, assuming these are basic goods—and yet be depressed about or even hate their reasoning or knowledge, it seems these goods are susceptible to this alienation objection. And so, we seem to have reason to doubt that they are basic (objective) goods.

Of course, it may be objected that, even if these things can for this reason not be basically good for *us*, it still might be that they are basically good for language agents. To admit this, though, would effectively be to give up the game. Goldstein and Kirk-Giannini are not, I take it, arguing that current language agents have "language-agent-well-being," where this is distinct from the kind of well-being you and I have. Instead, they're arguing that language agents have *well-being*—the same well-being as you and I. If they were not, then the implications of the leading theories of well-being—which concern you and I—wouldn't be relevant to their arguments. So, I take it, ruling out something as basically (or objectively) good for us is to rule out the thing as basically (or objectively) good for the language agents. Since we have reason to rule out rationality and knowledge as basically good for us, we have reason to rule out rationality and knowledge as basically good for the language agents.

[15] Railton 1986.

[16] Fletcher 2013.

[17] Notably, Fanciullo (2025) argues, more specifically, that only goods that require some phenomenologically salient, affective appeal on the part of the subjects possessing them can avoid this objection. And, if that's right, then given we've seen that existing language agents don't seem capable of such phenomenologically salient experience, we'll have reason to think these agents are incapable of well-being on *any* plausible version of the objective list approach.

2. Conclusion

In sum: the leading verisons of hedonism, desire satisfactionism, and the objective list approach do not seem to imply that current language agents are capable of welfare. Upon closer inspection, it seems the only route to the conclusion that current language agents are capable of welfare is through theories of well-being that we have strong independent reasons to reject. I conclude, therefore, that we have serious reason to doubt that current AI systems are capable of well-being.[18]

[18] Many thanks to Simon Goldstein, Dan Pallies, and two anonymous referees for helpful comments.

References

Butlin, P., R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, R. Kanai, C. Klein, G. Lindsay, M. Michel, L. Mudrik, M. A. K. Peters, E. Schwitzgebel, J. Simon, and R. VanRullen. (2023). "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." *arXiv preprint arXiv:2308.08708v3*.

Dorsey, D. (2010). "Three Arguments for Perfectionism." *Noûs* 44: 59-79.

Fanciullo, J. (2022). "On Sense and Preference." *Journal of Moral Philosophy* 19(3): 280-302.

— (2025). "Alienation, Engagement, and Welfare." *Philosophical Quarterly* 75: 40-60.

— (forthcoming). "Desire, Aversion, and Welfare." *Analysis*. doi: 10.1093/analys/anae101.

Fletcher, G. (2013). "A Fresh Start for the Objective-List Theory of Well-Being." *Utilitas* 25: 206-220.

— (2016). "Objective List Theories." In G. Fletcher (ed.), *The Routledge Handbook of the Philosophy of Well-Being*. New York: Routledge.

Goldstein, G. and C. D. Kirk-Giannini. (forthcoming). "AI Wellbeing." *Asian Journal of Philosophy*.

— (2025). "A Case for AI Consciousness: Language Agents and Global Workspace Theory." Unpublished manuscript.

Heathwood, C. (2016). "Desire-Fulfillment Theory." In G. Fletcher (ed.), *The Routledge Handbook of the Philosophy of Well-Being*. New York: Routledge.

— (2019). "Which Desires Are Relevant to Well-Being?" *Noûs* 53: 664-688.

Kitcher, P. (1999). "Essence and Perfection." *Ethics* 110: 59-83.

Levinstein, B. A. and D. A. Herrmann. (2024). "Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks." *Philosophical Studies*. doi: 10.1007/s11098-023-02094-3.

Pallies, D. (2022). "Attraction, Aversion, and Asymmetrical Desires." *Ethics* 132: 598-620.

Railton, P. (1986). "Facts and Values." *Philosophical Topics* 14: 5-31.

— (2012). "That Obscure Object, Desire." *Proceeds and Addresses of the American Philosophical Association* 86: 22-46.

Saad, B. and A. Bradley. (2022). "Digital Suffering: Why It's a Problem and How to Prevent It." *Inquiry*. doi: 10.1080/0020174X.2022.2144442.

Smithies, D. (forthcoming). "Hedonic Consciousness and Moral Status." In U. Kriegel (ed.), *Oxford Studies in Philosophy of Mind*, vol. 5. Oxford: Oxford University Press.