IRIE
International Review of Information Ethics Vol. 29 (03/2021)

Howard Nye and Tugba Yolbas

Artificial Moral Patients: Mentality, Intentionality, and Systematicity

Abstract

In this paper, we defend three claims about what it will take for an AI system to be a *basic moral patient* to whom we can owe duties of non-maleficence not to harm her and duties of beneficence to benefit her: (1) Moral patients are mental patients; (2) Mental patients are true intentional systems; and (3) True intentional systems are systematically flexible. We suggest that we should be particularly alert to the possibility of such systematically flexible true intentional systems developing in the areas of exploratory robots and artificial personal assistants. Finally, we argue that in light of our failure to respect the well-being of existing biological moral patients and worries about our limited resources, there are compelling moral reasons to treat artificial moral patiency as something to be avoided at least for now.

Keywords: Artificial Intelligence, Artificial Moral Patients, Mental Patients, Moral Patients, Well-being

Outline:

Authors:

Howard Nye

- University of Alberta, 2-59 Assiniboia Hall, 9137, 116 Street NW Edmonton, Alberta, T6G 2E7, Canada
- +1 780 492-8554, hnye@ualberta.ca

Tugba Yolbas

- University of Alberta, 2-26 Assiniboia Hall, 9137, 116 Street NW Edmonton, Alberta, T6G 2E7, Canada
- +1 780 707-2719, yoldas@ualberta.ca

ISSN 1614-1687

1. Introduction

Most work on ethics and artificial intelligence (AI) appropriately focuses on how designing and employing AI systems may affect individuals other than the AI systems themselves. We believe, however, that it is not premature to begin thinking about what it would take for an AI system to itself be a *moral patient*, to whom we would owe moral duties. The most basic moral duties we can owe to an individual are those of non-maleficence not to harm her, e.g., not to cause her to experience pain or suffering, or deprive her of future goods by killing her, and duties of beneficence to benefit her for her own sake, e.g. by alleviating her suffering and enabling her to enjoy her life[1]. Central to the question of whether an entity is a moral patient is thus whether she can be harmed and benefitted in a sense relevant to these duties.

In this paper, we defend three claims about what makes it the case that an AI system would be a *basic moral patient* to whom we can owe duties of non-maleficence and beneficence:

1. *Moral Patients Are Mental Patients*: an entity is a basic moral patient only if it is capable of mental states like experiences, motivations, and beliefs,

2. *Mental Patients Are True Intentional Systems*: an entity is capable of mental states only if it is a *true intentional system* that is best explained in terms of representations and goals, and

3. *True Intentional Systems are Systematically Flexible*: an entity is a true intentional system only if it is best explained as having representations and goals that can flexibly interact with a wide variety of other representations and goals in the way described by the philosophical theory of Success Semantics.

Because it is overwhelmingly unlikely that any present-day AI systems are systematically flexible true intentional systems, it is overwhelmingly unlikely that any are basic moral patients. This is good news, because we are doing a terrible job caring for biological moral patients, and the last thing we need are

[1] Any defensible moral view will need to acknowledge that at least part of morality is concerned with the well-being of individuals. Indeed, there is a great deal to be said in favour of what philosophers often call welfarism, or the idea that all of morality is essentially concerned with the welfare or the well-being of individuals, and that all of our moral duties can be understood as ultimately stemming from concerns to make individuals better off (see e.g. Keller 2009). This includes not only forms of consequentialism according to which morality is fundamentally concerned with trying to bring about the best outcomes understood in terms of the most well-being and the least ill-being of everyone (see e.g. Sidgwick 1907), but forms of pluralist deontology that follow W.D. Ross (1930) in holding that duties of non-maleficence are stronger than those of beneficence and that we have especially strong duties to do things that promote the well-being of those to whom we are specially related like our family members and those to whom we have made promises. But even if one is not a welfarist, and wishes to hold that there are, for instance, duties to respect the autonomy of competent adults that cannot be reduced to a concern for anyone's well-being, one must still acknowledge that basic duties of non-maleficence and beneficence that are concerned simply with well-being are at least as plausible, are needed to explain our duties to individuals like young children and those with severe intellectual disabilities, and are often needed to explain how we should prioritize the interests of competent adults. As such, even non-welfarists need to acknowledge that a basic *part* of morality concerns the well-being of moral patients in the form of basic duties of non-maleficence and beneficence. Moreover, even non-welfarists should acknowledge that it requires less mental sophistication for an entity to be capable of well-being than for her to have the kind of features that would make her capable of being owed more complicated duties like an irreducible respect for her autonomy. Thus, even non-welfarists should acknowledge that duties of non-maleficence and beneficence related to well-being are "most basic" in the sense that they require a most basic mental scaffolding to be owed. Finally, while we might have duties with respect to non-sentient nature, these duties are not, in the sense discussed below, owed *to* non-sentient entities in the same way that duties of non-maleficence and beneficence are owed to moral patients. Hence, we think there is a good sense in which any defensible moral view should be able to acknowledge that well-being-centered duties of non-maleficence and beneficence are the most basic duties we can owe to an individual.

artificial moral patients to whom we will also fail in our duties and whose adequate care would require the depletion of resources needed to care for biological moral patients. While we should be sensitive to the possibility of systematically flexible true intentional systems in many fields, the danger of such systems developing may be most acute in the domains of robots used for exploration and artificial personal assistants.

2. Moral Patients are Mental Patients

We can speak in some sense of things being good or bad for bacteria and cars, insofar as they interfere with their biological functioning, intended purpose, or structural integrity. But these are not harms or benefits in the same morally important sense as that in which suffering is a harm to us and enjoyment is a benefit to us. It would be indefensibly arbitrary to hold that an entity can be harmed and benefited in the same morally important sense as us but that we have no duty to omit to harm her or to benefit her on the mere grounds that she lacks certain intellectual abilities or biological features. Rather, our most plausible moral principles of non-maleficence and beneficence hold that, if an entity can be harmed or benefited in a morally relevant sense, then we have a duty of non-maleficence of at least some strength to her not to do what would harm her, and a duty of beneficence of at least some strength to her to do what would benefit her[2].

As such, whether an entity is a basic moral patient depends upon whether it can be harmed and benefited in a morally relevant sense. What it takes for an entity to be capable of being harmed and benefited in a morally relevant sense is examined by philosophical theories of well-being. These theories seek to establish what kinds of things can, on reflection, be defensibly held to be harmful and beneficial intrinsically, or in themselves and independent of their further effects. There are three main kinds of such theories:

A. *Experientialist Theories* hold that the only things that are intrinsically beneficial and harmful for individuals are phenomenally conscious states, or states there is something it is like to be in. One of the most famous forms of experientialism is hedonism, or the view that the only things that are intrinsically good and bad are, respectively, pleasure or enjoyment, and pain or suffering. But experientialists need not be hedonists; they can, for example, hold that ordinary experiences which are neither pleasant nor painful are intrinsically good as ways of experiencing the world, or states that are far more rich, vivid, and intense than states that there is nothing it is like to be in.

B. *Motivation Fulfillment Theories* hold that the only things that are intrinsically beneficial and harmful for individuals are those that fulfill certain of their intrinsic motivations, or things the individuals are motivated to bring about in themselves and independent of their further effects. Since individuals have intrinsic motivations to have or avoid having certain phenomenally conscious states (which is plausibly part of what makes them enjoyable or aversive), motivation fulfillment theories offer one account of what makes certain experiences intrinsically beneficial or harmful. But because individuals can also have intrinsic motivations to have and avoid having other things, like genuine as opposed to merely illusory friendships, knowledge, and achievement, motivation fulfillment theories can hold that these other things, which are not simply states of phenomenal consciousness, are also intrinsically beneficial and harmful to individuals.

[2] See, for instance, Singer 1975, DeGrazia 1996, 2016, McMahan 2002, Bostrom and Yudkowsky 2014, and Basl 2014. Basl has a similar argument that even if "benefits" like mere structural integrity are morally relevant, they are relatively trivial in comparison to those that involve mental states.

C. *Objective List Theories* hold that there is a plurality of things which are intrinsically beneficial and harmful to individuals - some of which are not simply phenomenally conscious states - the value of which is not determined simply by how strongly intrinsically motivated individuals are to have them. While objective list theories can take many forms, the most plausible versions hold that, in addition to good experiences, the intrinsic benefits include things like intimate relationships, knowledge, achievement, and virtue, while in addition to bad experiences the intrinsic harms include things like being oppressed, deception, failure, and vice. The most important thing for our purposes is that, on the most plausible objective list theories, while some intrinsic benefits and harms are not simply phenomenally conscious states, they essentially involve phenomenally conscious states or other mental states as parts. That is, one cannot have an intimate relationship with someone without having certain feelings and thoughts about her, one cannot be dominated by someone in the relevant sense without being less able than her to achieve one's desired outcomes, one cannot know anything without believing it, one cannot in the relevant sense achieve anything without trying or being motivated to bring it about, and so on. The view that mere states of our bodies could be in themselves beneficial or harmful to us without having any effect on our present or future mental states is deeply implausible[3].

Thus, whichever plausible theory of well-being we accept, an entity will be capable of being harmed and benefited in a morally relevant sense only if it is capable of mental states like experiences, motivations, and beliefs. Experientialist theories entail that an entity is capable of being harmed and benefited only if it is capable of the experiences that constitute intrinsic harms and benefits. Motivation fulfillment theories entail that an entity is capable of being harmed and benefited only if it is capable of having intrinsic aversions and motivations, the fulfillment of which constitutes intrinsic harms and benefits. Plausible objective list theories entail that an entity is capable of being harmed and benefited only if it is capable of the experiences and other mental states that are an essential part of the intrinsic benefits and harms on the list.

3. Mental Patients are True Intentional Systems

Drawing upon the ideas of plausible, non-arbitrary principles of non-maleficence and beneficence, and the most defensible philosophical theories of well-being, we have thus argued that for an entity to be a basic moral patient to whom we owe duties of non-maleficence not to harm it and beneficence to benefit it, the entity must be capable of having mental states like experiences, motivations, and beliefs.

What, however, does it take for an entity to be capable of these mental states? As several philosophers have argued, mental states are distinctive in that they have a kind of original intentionality, or way of representing or being directed towards the world that is capable of genuine error and unfulfillment[4]. As Fred Dretske (1994) has argued, while we can speak of tree rings representing or indicating the age of a tree, they are not in themselves incorrect or mistaken if they fail to correspond to the tree's age. By contrast, mental states like beliefs and perceptions represent the world in a way capable of being genuinely false or mistaken. Moreover, while items of public language like written and spoken sentences can represent the world in a way that can be genuinely mistaken, their ability to do so is derived from their expressing our mental states. By contrast, our beliefs' and perceptions' ability to represent the world in this way is original and underivative from simply being a conventional expression of some other state that can be genuinely mistaken.

Motivations like desires and urges do not "represent" the world in a strict sense, or have what following Elizabeth Anscombe (1957) is called a mind-to-world direction of fit, of trying (as it were) to fit

[3] On these theories of well-being see, for instance, Parfit 1984, DeGrazia 1996, and Hurka 2011.

[4] On underivative intentionality as the mark of the mental, see Brentano 1874, Chisholm 1957, and Dretske 1994.

the mind to match the world. But they do have a distinctive kind of intentionality that following Anscombe is called world-to-mind, of seeking (as it were) to fit the world to match the mind, or to bring about what they are motivations to bring about. These goals of states with world-to-mind intentionality can be fulfilled or unfulfilled in a way that is similar to that in which the representations of states with mind-to-world intentionality can be correct or incorrect. As states with world-to-mind intentionality, the sense in which desires and urges to bring about certain states of the world (including oneself) can be fulfilled or unfulfilled is genuine, and not simply a matter of their being part of a system that is doing or failing to do what an observer might expect it to do.

Most philosophers would agree that original intentionality capable of genuine error or unfulfillment (henceforth just 'intentionality') is the most philosophically puzzling feature of mental states like beliefs and motivations, which distinguishes them from physical states that are not also mental states (like, presumably, those of rocks and electrical currents). But it is not so obvious that intentionality is the most philosophically puzzling feature of subjective experiences or phenomenally conscious states. It is not even clear that all phenomenal states are intentional. While it is pretty uncontroversial that phenomenal states like perceptions represent the world, it is not as clear that experiences like itches, tingles, emotions, pleasures, and pains represent or are about the world - let alone that this intentionality is the main thing that makes them what they are.

There are, however, important philosophical arguments that intentionality is actually central to phenomenal consciousness. Perhaps the most prominent is what we might call the argument from transparency: that when we try to introspect and describe the essence of what it is like to have any experience, all we can find is what it represents, or what it involves an urge to do or bring about. This seems clearly to be true in the case of perceptions of the external world, but upon reflection, it also seems true of bodily sensations. As authors like Michael Tye (1995) convincingly argue, itches, tingles, and pains all seem to represent certain kinds of bodily disturbances - which representations can in fact be mistaken in the case of such things as phantom limb phenomena where one may experience sensations that represent disturbances in a limb that has been removed. Similarly, states like emotions and sensations of feeling hot or cold seem to represent more global bodily changes of autonomic arousal or increased or decreased body temperature.

While all phenomenal states may have representational aspects, these may not seem to exhaust their distinctive phenomenal features. For example, the sensory aspects of pain can be dissociated from pain's affective dimensions, and while a sensory pain that does not hurt might have the same sensory aspects as a pain that does, the pain that hurts seems to feel differently in an important non-sensory way. Moreover, when we introspect and seek to explain what it is like to experience things like itches, tingles, feelings of hunger or thirst, and emotions, we seem to find features other than the bodily disturbances or changes that they represent. Part of what it is like to experience these things seems to involve urges to scratch, eat, drink, escape, or retaliate (Hall 2008). The way they feel seems also to involve their being either pleasant or aversive, such that we like or dislike them, or such that we have an inclination for them to continue or stop. But while these aspects of phenomenal consciousness are not representational or such as to involve mind-to-world intentionality, they do seem to involve motivational goals, or constitute a form of world-to-mind intentionality.

It seems, then, that a key part of what it is for an entity to be capable not only of mental states like beliefs and desires but indeed of phenomenally conscious states is for it to have states with intentionality. What, however, does it take for an entity to have intentional states?

As Daniel Dennett (1981) argues, we can, if we like, take the "intentional stance," or attribute intentional states like beliefs and desires to a wide variety of systems when we describe their behaviour. We can, for instance, describe an ordinary thermostat as wanting the temperature of a room to be no higher than the set-point, and acting so as to heat or omit heating the room depending upon whether it believes the temperature is or is not above the set-point. But as Dennett argues, there is a way in which attributing intentional states to entities like us is an indispensable part of the best explanation of our

behaviour, but not that of entities like ordinary thermostats. As Dennett (1981, 1991) argues, attributions of intentional states to "true believers" or "true intentional systems" like us pick out a "real pattern" or explanatorily important regularity in our behavioural dispositions, which is not the case of attributions of such states to merely metaphorical believers or intentional systems like ordinary thermostats. The real pattern as Dennett describes it is roughly that we have representational and motivational states that are disposed to combine with a wide variety of other such states in a characteristic way to explain a wide variety of different behaviours.

4. True Intentional Systems are Systematically Flexible

As Dennett (1971) argues, there are some contexts in which present-day AI systems can seem in a sense to be intentional systems. There is a sense in which attributing the goal of winning a chess game, various intermediate goals, and various representations of what will achieve those goals may pick out a kind of real pattern in explaining the behavior of a chess-playing AI. We believe, however, that there is an important distinction between these systems and those Dennett (1981) calls "true believers" or true intentional systems.

In characterizing the chess-playing AI as having the goal of winning at chess, or representations of what will, say, cause it to capture its opponent's pieces, we are clearly attributing to it goals and representations that are radically unlike our goals of winning at chess and representations of what will enable us to capture our opponent's pieces. Our goals and representations involve elements corresponding to winning, games, pieces, movement, capturing, and so on that can be used in many contexts outside of playing chess. This is part of what enables us to literally characterize each other as having goals and representations of winning, pieces, movement, and so on: plausibly, part of what makes a state of me a representation of a piece being moved is that it has elements that are disposed to combine with different representations and goals in different contexts related to pieces and movement to cause very different behaviours quite outside of chess. Because the chess-playing AI only exhibits real patterns relevant to the playing of chess, there is nothing that allows the states of these patterns to be interpretable as having truly determinate representational contents. We simply describe its states as those about chess, pieces, and winning because those are the purposes with which we have designed it, or which we have in interacting with it.

What does it take, then, for an entity to exhibit the kind of systematic flexibility that seems to be required for it to have states with the kind of determinate content required for it to constitute a true intentional system? We believe that the most promising general answer is given by Success Semantics, as developed by J.T. Whyte (1990, 1991). The basic idea of success semantics is that we can explain the contents of a system's states in terms of what would successfully achieve the different possible goals it could have if it were to act in light of the different possible representations it could have. In a little more detail, the first idea of success semantics is roughly that:

(R) What it is for a state of a system, *B*, to be a representation that *P* is for *B* be disposed to combine with the system's goals to cause behavior that would achieve those goals if *P*.

For example, what makes it the case that some (e.g. neural) state of Mary is a representation that there is a desk in her room is that this state is disposed to combine with her goals to cause behaviour that would achieve those goals if there were a desk in her room. The state would, for instance, combine with her goal for the room to contain exactly one desk to cause her to reject her friend's offer of another desk.

Of course, what a representation and goal cause a system to do depends upon its other representations. J.T. Whyte argues that, to be precise, success semantics should begin by giving an account of what it is for a set of states of an agent to be a set of representations that certain things are true in

terms of the conditions under which the actions motivated by that whole set of representations together with the agent's goals would fulfill its goals. Success semantics should then go on to explain what it is for a state to be a representation of a particular state of affairs in terms of what is common to the conditions under which the actions motivated by the various sets of representations in which it could be a member, together with the agent's goals, would actually achieve those goals. We can also explain what it is for a smaller component state (like our representation of the property of being moved) to be a representation of a component of a state of affairs (like the property of being moved) in terms of what is common to the conditions under which all of the states in which it occurs would achieve the goals with which they were combining.

Since R explains what it is for a state to be a representation in terms of its interaction with goals for certain things, success semantics must also provide an account of what it is for a state to be a goal for something. Some goals seem explicable only in terms of the system's representations in the following way:

(F) What it is for a state of a system, *G*, to be a goal that *O* (i.e. that will be fulfilled by *O*) is for *G* to be disposed to combine with the system's representations to cause behavior that would bring it about that *O* if those representations were accurate (i.e. if their contents were to obtain).

For instance, the causal essence of a goal to have exactly one desk seems to be a disposition to combine with one's representations (e.g. that one already has a desk) to cause one to do things (like decline an offer of another desk) that would bring it about that one has exactly one desk if those representations were accurate (e.g. one already has a desk).

A theory of true intentionality composed of R and F alone might not seem very informative; we might worry that its explanation of representations and goals in terms of each other would say too little about their natures in independent terms, making it viciously circular. But as Whyte (1991) argues, we should hold that there is a set of "basic" goals, the contents of which can be explained without reference to the content of any of the agent's representations. With this characterization of basic goals in hand, we can use R and F to recursively build upon its foundation to give an account of the contents of representations and other goals. Whyte's proposal is that we can understand these basic goals in terms of what "reinforcingly satisfies" them, or:

(S) What it is for a state of a system, *G*, to be a basic goal that *O* is for it to be the case that *O* would (i) cause *G* to "go away" or cease exerting causal influence on the system's behaviour, (ii) in a way that would reinforce the system's disposition to act in the way that led to *O* when *G* is next present or active.

For instance, an important part of the basic goal of tasting something that tastes like cherries involved in having an inclination to taste cherries is that getting this taste causes the goal to go away or cease influencing one's conduct. Of course, a basic goal can cease influencing one's conduct because of things other than its being achieved, such as one's receiving a hard blow to the stomach. But when one lacks a goal of receiving a blow to the stomach, one's receiving one does not tend to cause one to repeat whatever led to one's receiving the blow. On the other hand, it seems plausibly essential to a basic goal for something like tasting cherries that one has a tendency to become more likely to do whatever it was that led to one's tasting them if one is motivated by the goal in the future.

5. Conclusion

We have thus argued that for an entity to be a moral patient it must have mental states like experiences, motivations, and beliefs, and that in order to have such mental states it must be a systematically flexible

true intentional system with states that count as representations and goals as characterized by the success semantical principles R, F, and S. There are excellent reasons to believe that at least all vertebrate animals, cephalopod molluscs, and big-brained arthropods like crustaceans, bees, and spiders are systematically flexible true intentional systems. These individuals have representations and goals that are poised to flexibly cause different kinds of behaviour depending upon the other representations and goals that they come to have[5]. But because present-day AI systems so far have relatively narrow task ranges, it seems extremely unlikely that any have states that exhibit the kind of systemically flexible true intentionality characterized by principles R, F, and S.

By any reasonable estimation, we are doing a terrible job respecting the well-being of the clear biological moral patients who inhabit the earth. The insane human practice of needlessly consuming animal flesh and secretions, and consuming extreme amounts of them[6], is continuing to destroy the environment through humans using far more land and water and generating far more polluting waste than they need, simply to breed tens of billions of land animals and hundreds of billions of farmed fish, confine them, cycle energy through them, kill them, and get a tiny amount of energy back from their secretions and corpses in one of the most radically inefficient process of obtaining nutrients imaginable. This ecological catastrophe, together with humanity's continued failure to switch to a decarbonized energy system, is also causing a climate emergency that threatens to wipe out hundreds of millions of humans, and orders of magnitude more wild animals[7].

As such, the very last thing we should do is create an entire new set of moral patients in the form of artificial moral patients. Our track record suggests that we will treat them just as abysmally as the biological moral patients we are routinely destroying. Any attempt to treat them decently would also require the depletion of resources needed to mitigate the vast harms we are doing to the already existent biological moral patients. The mere fact that a possible moral patient who does not yet exist would experience more 'goods' than 'bads' if we were to create her is, moreover, no reason to actually create her; we have no reason, for instance, to try to create every possible human who would experience more 'goods' than 'bads' that we can. As authors like Melinda Roberts (2011) have argued, existence is a precondition for a moral patient to be harmed in a morally important sense, in that her existence in a scenario is necessary condition for the fact that she could have had more well-being in another scenario to count in favour of bringing about the alternative scenario for her sake.

Thus, while AI systems can have important instrumental benefits to us, we should at least for now treat artificial moral patiency as something to be avoided. If we are correct that a key necessary condition for artificial moral patiency is systematically flexible true intentionality, then it seems likely that the greatest risks of artificial moral patiency are with AI systems that have flexible behavioural repertoires across a wide variety of domains. This might include robots used for exploration like the Mars Curiosity Rover and automated personal assistants, especially those that can act across a wide variety of domains in contexts like integration with automated or smart homes. While we should be on guard to prevent the emergence of artificial moral patiency in many domains, the domains of robots used for exploration and personal assistants may be the ones in which we should most actively seek to distance our AI systems from having the kind of systematically flexible true intentionality that we see most clearly instantiated in non-human animals like vertebrates, cephalopods, and big-brained arthropods.

[5] See Papini 2008, Panksepp 2005, Braithwaite 2010; Mather 2008, Tye 1997, and Elwood 2011.

[6] Because plant-based diets are at least as healthy and have health benefits in relation to diets containing animal products, all consumption of animal products is needless (see Academy of Nutrition and Dietetics 2016), and the extreme amount of animal products consumed is both needless and ecologically catastrophic.

[7] See Oppenlander 2013, Poore and Nemecek 2018, Bar-On et al., 2018, Shepon et al. 2018, Searchinger et al. 2018, Nolt 2011, IPCC 2014, Tomasik 2019.

6. References

Academy of Nutrition and Dietetics. "Position of the Academy of Nutrition and Dietetics: Vegetarian diets." Journal of the Academy of Nutrition and Dietetics, vol. 116, 2016, pp. 1970-1980, https://doi.org/10.1016/j.jand.2016.09.025.

Anscombe, Gertraude E. Intention. Blackwell, 1957.

Bar-On, Yinon M., Phillips Rob, and Milo Ron. "The Biomass Distribution on Earth". Proceedings of the National Academy of Sciences of the United States of America, vol. 115, no. 25, 2018, pp. 6506-6511. DOI: 10.1073/pnas.1711842115.

Basl, John. "Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines". Philosophy and Technology, vol. 27, no. 1, 2014, pp. 79-96.

Bostrom, Nick and Yudkowsky, Eliezer. "The Ethics of Artificial Intelligence". In Frankish, Keith and Ramsey, William M. (eds), Cambridge Handbook of Artificial Intelligence, Cambridge University Press, 2014, pp. 316-334.

Braithwaite, Victoria. Do Fish Feel Pain? Oxford University Press, 2010.

Brentano, Franz. "The Distinction Between Mental and Physical Phenomena". 1874. In Franz Brentano, Psychology from an Empirical Standpoint, translated by A.C. Rancurello, D.B. Terrell, and L. McAlister, London Routledge, 1973 (2nd ed., intr. by Peter Simons, 1995).

Chisholm, Roderick. Perceiving: A Philosophical Study. Cornell University Press, 1957.

Degrazia, David. Taking Animals Seriously: Mental Life and Moral Status. Cambridge University Press, 1996.

Degrazia, David. "Modal Personhood and Moral Status: A Reply to Kagan's Proposal". Journal of Applied Philosophy, vol. 33, no. 1, 2016, pp. 22-5. DOI: 10.1111/japp.12166

Dennett, Daniel. "Intentional Systems". Journal of Philosophy, vol. 68, no. 4, 1971, pp. 87-106.

Dennett, Daniel. "True Believers: The Intentional Strategy and Why It Works". In A. F. Heath (ed.), Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford, Clarendon Press, pp. 150-167. Reprinted in Chalmers, David (ed.), Philosophy of Mind: Classical and Contemporary Readings, Oxford University Press, pp. 556-568

Dennett, Daniel. "Real Patterns". Journal of Philosophy, vol. 88, no. 1, 1991, pp. 27-51.

Dretske, Fred. "If You Can't Make One, You Don't Know How It Works". Midwest Studies in Philosophy, vo. 19, no. 1, 1994, pp. 468-482.

Elwood, Robert W. "Pain and Suffering in Invertebrates". ILAR [Institute for Laboratory Animal Research] Journal, vol. 52, no. 2, 2011, pp. 175-184.

Hall, Richard. "If it Itches, Scratch!" Australasian Journal of Philosophy, vol. 86, no. 4, 2008, pp. 525–535.

Hurka, Thomas. The Best Things in Life: A Guide to What Really Matters. Oxford University Press, 2011.

IPCC 2014, Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, ed. by Core Writing Team, R.K. Pachauri, and L.A. Meyer (IPCC: Geneva, Switzerland, 2014)

Keller, Simon. "Welfarism". Philosophy Compass, vol. 4, no. 1, 2009, pp. 82-95.

Mather, Jennifer. "Cephalopod Consciousness: Behavioural Evidence". Consciousness and Cognition, vol. 17, 2008, pp. 37-48.

McMahan, Jeff. The Ethics of Killing: Problems at the Margins of Life. Oxford University Press, 2002.

Nolt, John. "How Harmful Are the Average American's Greenhouse Gas Emissions?" Ethics, Policy and Environment, vol. 14, 2011, pp. 3-10.

Oppenlander, Richard. Food Choice and Sustainability: Why Buying Local, Eating Less Meat, and Taking Baby Steps Won't Work. Langdon Street Press, 2013.

Panksepp, Jaak. "Affective Consciousness: Core Emotional Feelings in Animals and Humans". Consciousness and Cognition, vol. 14, 2005, pp. 30-80.

Papini, Mauricio R. Comparative Psychology: Evolution and Development of Behavior, 2nd Edition. Psychology Press, 2008.

Parfit, Derek. Reasons and Persons. Oxford University Press, 1984.

Poore, Joseph and Nemecek, Thomas. "Reducing Food's Environmental Impacts Through Producers & Consumers". Science, vol. 360, 2018, pp. 987-992. DOI: 10.1126/science.aaq0216.

Roberts, Melinda. "The Asymmetry: A solution". Theoria, vol. 77, 2011, pp. 333-67.

Ross, W. D.. The Right and the Good. Clarendon Press, 1930.

Searchinger, Timothy Stefan Wirsenius, Tim Beringer, and Patrice Dumas. "Assessing the Efficiency of Changes in Land Use for Mitigating Climate Change". Nature, vol. 564, 2018, pp. 249-253, https://doi.org/10.1038/s41586-018-0757-z.

Shepon, Alon, Gideon Eshel, Elon Noor, and Ron Milo. "The Opportunity Cost of Animal Based Diets Exceeds All Food Losses". Proceedings of the National Academy of Science [PNAS], vol. 115(15), 2018, pp. 3804-3809 , https://doi.org/10.1073/pnas.1713820115.

Sidgwick, Henry. The Methods of Ethics. 7th Edition edn. Macmillan and Co., Limited, 1907.

Singer, Peter. Animal Liberation. Harper Collins, 1975.

Tomasik, Brian. "How Many Wild Animals Are There?". 7 Aug. 2019. http://reducing-suffering.org/how-many-wild-animals-are-there/. Accessed 10 January 2019.

Tye, Michael. Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind. MIT Press, 1995.

Tye, Michael. "The Problem of Simple Minds: Is there Anything it is Like to Be a Honey Bee?". Philosophical Studies, vol. 88, 1997, pp. 289-317.

Whyte, Jamie T. "Success Semantics". Analysis, vol. 50, no. 3, 1990, pp. 149-157.

Whyte, Jamie T. "The Normal Rewards of Success". Analysis, vol. 51, no. 2, 1991, pp. 65-73.