Fichte on the Value of Rational Agency

What role does the idea of the value of humanity play in Fichte's ethics? It seems natural, at least initially, to think it must play a very central role. Fichte saw himself as a Kantian in ethics; and it is after all Kant's name that the phrase 'value of humanity' calls first to mind. But this quite natural thought would have to be reconciled with the fact that the phrase 'value of humanity' occurs only once in Fichte's published works, and that its components appear only seldom, and not at all in the ways one might expect (§1).[1] Such a reconciliation would not be impossible: Kant's use of 'humanity' is technical, and Fichte's theory is (like Kant's) one in which rational nature is an obligatory end (§2). But to think of Fichte's theory in terms of the value of humanity may in the end be undesirable, since on it the features in virtue of which agents are worthy of moral consideration neither depend on nor are guaranteed by membership in the human species (§3). So in contexts that do not limit the reference of 'humanity' in the technical sense to human beings in the biological sense, the use of that term is, at best, misleading (§4).

§1. Texts

The term 'value' appears infrequently in Fichte's ethical writings, and strikingly not at all in the first, foundational division of his main work of ethical theory, the 1798 *System of Ethics*. In fact the term appears only 16 times in the entire work, and only in the third division, in the sections devoted to the application of the ethical principle. In those occurrences where the value of human beings in particular is at issue (as opposed to the value of goods or types of labor), the context is typically a denial that differential moral regard for individuals is legitimated by their different places in some status hierarchy. These passages express Fichte's commitment to the equal claim to moral consideration of all beings with any claim at all, regardless of their socioeconomic class or caste, their degree of moral virtue, or other differences.[2]

This lack of instances of 'value' in the expected meaning is unsurprising: like Kant's, Fichte's moral theory is not one in which an account of value has a foundational role. Foundational, instead, is an account of what rational agents will insofar as they are rational. On Fichte's account, they will certain ends constitutive of rational agency; and this is, plausibly, why Fichte prefers talk of ends to talk of values. (The term 'end' and its derivatives occurs over 300 times in the *System of Ethics*, and occurs throughout the work.) One might legitimately

[1] The single passage in which the phrase 'value of humanity' occurs in Fichte's published work is one in which he explains that the value of humanity consists of its exceptional exemplars, those who have, through great effort, developed their human potential to the highest degree (V: 426). Clearly, this is far from the usual employment of the phrase in contemporary Kantian ethics. I discuss the components further in §1.

[2] Fichte's target in these passages appears to be theories, like Godwin's, on which certain sorts of people (e.g. Fenelon) are more worthy of moral consideration because of intrinsic properties than other sorts of people (e.g. Fenelon's chambermaid).

describe those ends as having 'objective value,' but in doing so one would have to bear in mind that the only thing that can mean, in the Fichtean context, is that their pursuit is rationally obligatory. It cannot mean that their value is something independent that grounds the rationality of their pursuit. On this count, though, there is no disagreement with Kant, for whom talk about humanity having 'value' is just a variant of talk about humanity being an 'objective end' or an 'end in itself'.

The term 'humanity' appears only 21 times in the *System of Ethics*, and the pattern described above concerning 'value' is repeated: no occurrences at all in the deduction, and all but one occurrence in the sections devoted to application. Fichte uses other terms to refer to moral agents, chiefly 'rational being'; and he never uses 'humanity' as Kant does, to refer to a (special morally relevant) *property* of human beings (a sense of 'humanity' in which it makes sense to talk about humanity '*in* oneself or others'). Nor does he describe humanity as being an object of respect or as possessing a special dignity or as being an objective end or an end in itself. Strikingly, the one occurrence of the term in which it designates, in Kantian fashion, that feature in virtue of which human beings merit moral respect, is within a quotation from Schelling at IV: 225. In his mature works Fichte uses 'humanity' exclusively as a mass noun to refer to the human species (for instance, to describe humanity's progress over history (as at IV: 240, 241, 256) or to describe something as good or useful for human beings generally (as at IV: 176 or VI: 328–29)).[3]

This is also in the end not very surprising, since on Fichte's theory, human beings simply as such have no particular value and are in no sense the end at which moral action aims. That end is the exercise of rational agency; and although there is a meaningful overlap between the set of rational agents and the members of the human species, the overlap is far from complete. This also seems to me to reflect no fundamental disagreement with Kant, who after all tells us that what he means by 'humanity', in the passages in which he describes it as an end in itself, is *rational nature*, and who discusses, explicitly, the possibility of non-human finite rational agents no less bound by the moral law than we are.[4] Still, there is something more radical about Fichte's divorce of rational nature from the biologically human than most interpreters seem to see in Kant.

Before explaining why (§3) I will give a brief overview of Fichte's ethical theory (§2), because it will be unfamiliar to most readers. (The next section draws heavily on some recent work of mine.[5] Readers familiar with that work may want to skip it.)

§2. Fichte's ethical theory

[3] The only work containing language echoing the Kantian use of 'humanity' is the 1792 *Attempt at a Critique of all Revelation*, in which Fichte follows Kant on a number of other issues as well on which he will almost immediately after its publication begin to depart from Kant.
[4] I. Kant 1900- 4:408; I Kant 1996 p. 62.
[5] M. Kosch 2014, 2015a, 2017 and 2018.

Fichte's theory is a form of constitutivism on which rational agency has a necessary end, and moral obligations are construed as rational obligations to act in ways that further that end. The end is independence (alternatively, self-sufficiency); and independence has a formal and a material aspect.

Formally, independence provides a standard of correctness of *deliberation*. An agent is not deciding (fully formally) rationally unless she is doing so conscientiously in a double sense: letting her own conviction be her guide about what the right thing to do is in the situation (instead of deferring to someone else's opinion); and making sure she has a (genuine) conviction to act on (instead of acting thoughtlessly or impulsively). When an agent performs the action that is dictated by the conviction that issues from the sufficiently energetic application of her reflective capacities in a given situation, and performs it because it is so dictated, she acts independently in this formal sense.[6]

Materially or substantively, independence provides an objective standard of correctness of *actions*: An agent is not acting (fully materially) rationally unless she is taking the best means at her disposal to remove or decrease substantive limitations along one (or more) of the dimensions essential to rational agents. Fichte derives an account of these dimensions in an exercise of transcendental anthropology in the 'deduction of principle' sections of the two major works of practical philosophy of the Jena period (the 1796–97 *Foundations of Natural Right* and the 1798 *System of Ethics*). The idea is to begin from the premise that reflective self-consciousness is possible (a premise Fichte assumes his reader will grant) and to argue for these dimensions of limitation as necessary conditions of self-consciousness.

The conclusion of this set of arguments is that any intellect conscious of itself must be *practical* (a willing, not merely a thinking, being), *embodied* (having real, if finite, external causal powers), and *one among many* such beings (an 'individual'). Duties concerning agents as practical reasoners, as embodied physical causes, and as inhabiting a shared world with other such agents, follow from this division. Moral duties are, then, duties to overcome obstacles to the exercise of the practical intellect, to the execution of rationally formed plans, and to the coordination of individual activity in pursuit of these first two ends.[7] (Fichte's normative theory is thus consequentialist in structure,[8] although his

[6] See M. Kosch 2014 for a fuller treatment of conscience in Fichte's theory.
[7] See M. Kosch 2015a and 2018 for a fuller treatment of the substantive condition on moral worth in Fichte's theory.
[8] Fichte embraces a maximizing, fully agent-neutral form of consequentialism. See M. Kosch 2018, chapter 3, for a fuller explanation.

account of foundations is Kantian,[9] and although it incorporates an independent formal condition on correct deliberation.[10])

A priori principles governing the coordination of the activity of rational agents who are individuals are treated in the *Foundations of Natural Right*. Although Fichte takes some principles governing rational coordination to be a priori, he takes all actual rights to be the products of actual schemes of coordination and so to be entirely conventional. The purpose of such schemes is to allow rational agents who are multiple to 'coexist as free', by which Fichte means: to exercise their causality in the shared external world without inevitably undermining others 'efforts to do the same. To do this, they must divide up the space of *possible* action into non-overlapping individual spheres of *permissible* action, and organize themselves politically so as to guarantee the inviolability of these spheres.

So rational agents need to together develop a structure of rules and roles that order their collective existence, and the result is not only the foundation but also a good part of the edifice of their collective moral life. Although Fichte is a kind of act consequentialist, he believes that virtually everything rational agents accomplish (including what might appear to be solitary pursuits) relies on interpersonal cooperation, and so he also believes that the question of what an individual ought to do in a given situation is almost always settled by the conjunction of the laws they live under and the particular ends and duties that go along with their place in the social division of labor.

For the same reason, he denies that there is a morally sanctioned way of interacting with adults who, for whatever reason, refuse to become or remain members in good standing of some political community. Full moral standing, on his theory, relies on active membership in some such community.

---

[9] One finds this kind of theory structure in R. M. Hare 1997, or in D. Cummiskey 1996, for example. What we have reason to do is cashed out in terms of what it is rational to do; but what it is rational to do involves, in part, acting toward certain ends. So Fichte's theory is teleological in one of the two ways in which that term is used in contemporary ethical theory: its moral principle judges acts correct just in case they promote a specified end. It is non-teleological in the other current sense: it does not derive claims about what one ought to do from independent claims about what is good. It is, to use Cummiskey's language, a normative consequentialism without being a foundational consequentialism.

[10] The two criteria of evaluation are independent because an action can be formally correct (that is, conscientiously undertaken) without being, objectively, the action in the circumstances most conducive to the moral end; or an agent may rashly (and so unconscientiously) perform an action that is objectively the correct one in the circumstances. That independence in the formal sense is a constitutive end of rational agency is of course a substantive claim. One might think that, given that he recognizes an objective criterion of right action (*viz*. material independence) Fichte ought to say that the right way to make decisions is just whatever way turns out to best further that end (and for all we know that might be acting on instinct, or on the rules learned in the nursery). But Fichte disagrees, both because he thinks that instincts and rules learned in the nursery often lead us substantively astray, and because he thinks we have an independent interest in the rationality, deliberateness, independence of our procedure in deciding what to do.

§3. The extent of the moral community

The set of beings owed moral consideration is, then, the set of those integrated (or, with caveats I will shortly explain, integratable) into some political community.[11] The problem of identifying (potential) co-citizens is therefore not only politically but also morally extremely important, as Fichte explains in this passage from the *Foundations of Natural Right* (whose context is a discussion of who should be considered potential fellow citizens):

> Kant says: act in such a way that the maxim of your will can be a principle of universal legislation. But who belongs to the kingdom ruled by this legislation, and who is entitled to its protection? I am supposed to treat (*behandeln*) certain beings in such a way that I can will that they treat me in turn according to the same maxim. But I act (*handele*) every day upon animals and inanimate objects without ever seriously raising that question. Someone may say: obviously, only beings capable of the representation of a law, thus rational beings, are at issue. But this only substitutes one indeterminate concept for another, and does not answer my question at all. For how do I know which determinate object is a rational being? (III: 80–81)

He goes on to ask, in particular, about the status of children, domestic animals, and non-Europeans. He does not ask about women in this context, but his later arguments against according them full legal personhood (III: 304–312) make them another problematic case.

Fichte's discussion here recalls Bentham's (roughly contemporaneous) remarks about the moral status of entities that are not adult men of European ancestry in the note to chapter 17 of the *Introduction to the Principles of Morals and Legislation*.[12] Both Fichte and Bentham are writing at a time when slavery is still practiced in the Americas and the transatlantic slave trade is still tolerated by the international community, when women have the full legal rights and responsibilities of men nowhere in the world, and when the consensus concerning the moral standing of children and animals is that they lack any. Both see the question of moral standing as a pressing one, given these facts.

For Bentham, the commonality defining morally considerable beings is sentience; so he takes the extent of the set of morally considerable beings to be (relatively) easily fixed. For Fichte the question is more vexed: he takes susceptibility to pleasure and pain to be irrelevant to moral standing (though it may be indirectly relevant to some moral duties); but he has no similarly straightforward criterion to substitute for it.

---

[11] What I say in this section is an interpretation of Fichte's views during the Jena period. His 1813 lectures on political philosophy (*Die Staatslehre, oder über das Verhältniss des Urstaates zum Vernunftreiche*) seem to me to present a view in some respects irreconcilable with this, and so I have not tried to integrate them here.

[12] J. Bentham 1907 pp. 310–11.

This might seem strange: has he not, in articulating the theory of agency on which his practical philosophy depends, already provided the relevant criterion? The moral community is made up of those beings who are self-conscious and therefore have the essential features of self-conscious beings (are embodied practical intellects that are individuals). But in the above passage he seems to be claiming that this answer to the question does not suffice.

On reflection we can see that this is true. Part of what it is to say that rational agents are individuals is to say that they lack direct access to one another's mental lives. So to reply that those entitled to the protection of the moral law are the beings who are 'rational' and 'capable of the representation of a law' does not answer the question, in the absence of some behavioral criterion for determining which beings those are. That criterion would then do the actual work of sorting the potential co-citizens from the rest. But delineating such a criterion is far more problematic than delineating a behavioral criterion for sentience, for two reasons.

The first arises from Fichte's view, argued in §§3–4 of the *Foundations of Natural Right*, that a kind of socialization is a necessary condition of reflective self-consciousness. Reflective self-consciousness is the disposition to reflect on one's first-order attitudes and one's own behavior, to form higher-order attitudes, and eventually to adopt principles governing belief and action. So reflective self-consciousness is necessary for rational agency and its attendant capacities. If its existence relies on socialization, so too does rationality, and with it the capacity to represent a law. The socialization required for it is practice in the form of social interaction Fichte calls 'free reciprocal activity'. (I argue elsewhere that free reciprocal activity is cooperation, in any of its many forms.[13]) The invitation to such activity Fichte calls a 'summons' or, alternatively, 'upbringing'. So an unsocialized animal, no matter what its intellectual capacities and no matter what its interests, would not become reflectively self-conscious, and its rational potential would never be actualized. If the question Fichte is asking in the above passage is: who is the appropriate *target* of such socialization (as it plausibly is), the response cannot refer to behavior which is, by hypothesis, impossible prior to such socialization. So the absence of the behavior characteristic of reflectively self-conscious adults cannot indicate the absence of rational potential. (Sentience, by contrast, is not essentially reliant on socialization.)

The second reason for the difficulty of articulating a behavioral criterion is that, on Fichte's view, even achieved reflective self-consciousness cannot display itself as such in the wrong social conditions. That is because the behavior that would display it — free reciprocal activity — requires, as its name suggests, another rational agent simultaneously engaged in it. And recognition of something as a rational being is impossible in the absence of such behavior. So two (or more) agents become recognizable to one another as rational beings only if both (or all) engage in free reciprocal activity with one another, as Fichte here explains:

[13] I further explore the connection between cooperative interactions and the possibility of reflective agency in M. Kosch 2020.

> The relation of free beings to one another is therefore necessarily to be understood in the following way, and is posited as being so determined: the knowledge of the one individual by the other is conditioned on the other's treatment of it as free (that is, that the other limit his freedom through the concept of the freedom of the first). This mode of treatment is however conditioned on the action of the first toward the second, this action through the action and through the knowledge of the second, and so on to infinity. The relation of free beings to one another is thus the relation of interaction through intelligence and freedom. Neither can recognize the other if both do not reciprocally recognize one another. And neither can treat the other as a free being if both do not reciprocally treat one another that way. (III: 44)

The minimal cooperative scheme on Fichte's picture is mutual noninterference: abstaining from physical assault on one another's bodies, and dividing the space of possible activity in a way that allows each to pursue some set of activities un-interfered with by the other. The cooperative schemes characteristic of social life are more complex. They include shared language, customs, and norms, which afford ways of displaying rationality even to third parties with whom one is not currently *inter*acting. But the availability of these ways does not undercut the condition that Fichte here articulates, and so does not preclude the possibility of a normally socialized adult able to engage in this sort of behavior but not given the opportunity to engage in it being thereby rendered unable to present to others as a rational being. A captive in chains, deprived of the ability to act in a way that demonstrates that she grasps the freedom and rationality of her captors, would also be deprived of the opportunity to demonstrate her rationality to them.[14]

These considerations explain why simple appeal to the theory of rational agency cannot answer the question posed in the passage, which we might now rephrase in this way: How can agents living in a world full of items many of which are causally efficacious, which act in a way that is apparently purposive, and which display some intelligence, sort these into the ones that are potential fellow citizens, and the ones that are not?

Fichte offers, in the *Foundations*, what look like three different answers to this question.

---

[14] Note that the point concerns only the (self-imposed, in this example) epistemic limitations of the captors. Fichte's claim, in the discussion of cosmopolitan right, that there is a moral obligation to allow foreign visitors to 'explain themselves' (III: 384) is an application of this thought. They must be engaged, however briefly, in the free reciprocal activity that is the offering and consideration of 'transactions' (which might range from economic or other exchange to application for citizenship), which is equivalent to giving them the opportunity to demonstrate their status as rational beings and potential collaborators. Similar considerations support allowing growing children age-appropriate freedoms (III: 358–59).

His reasoning in the second division seems to conclude in the position that nature settles this question for us by giving (potentially) rational agents bodies of a certain sort. There are actually two accounts there, not clearly distinguished.

On the more prominent one (III: 78–80, 81–82), the form of the body is, in the human case, a kind of natural sign of its rational potential. Human bodies are distinguished from non-human animal bodies in *lacking an articulation* that defines a determinate sphere of movement, having instead an articulation that permits 'infinite determinability' (III: 79).[15] On a second account (only suggested, not explicitly spelled out, at III: 80–81), we recognize other rational beings by the *similarity* of their bodies to ours. If this were indeed Fichte's view, 'humanity 'in the sense of membership in the human species would have real moral significance in his theory. The similarity condition faces problems, of course, insofar as similarity admits of degrees, and the groups about which Fichte is asking are all to some extent similar to, and to some extent dissimilar to, adult European males. But in any event Fichte does not develop the physical similarity criterion, and may not even mean to suggest it in these brief remarks.

Later in the *Foundations* we find a third account, on which we recognize other rational beings by certain features, not of their embodiment, but of their behavior in interaction with us. On this account, it seems that we learn who the potential other rational agents are by *trial and error*. So for instance in §43 of the first appendix Fichte writes:

> It is a natural drive in human beings to suspect beings outside of themselves of rationality, where this is at all plausible, and to treat objects (for example, animals) as though they had it. The parents will treat their child in the same way, summoning it to free activity; and in this way rationality and freedom will gradually become manifest in it. (III: 358)

In the Hoijer Nachschrift of the 1798 lectures on logic and metaphysics we find a similar thought. Fichte points out that in early human history people attributed rational agency to inanimate objects like rivers, and he mentions the example of animals in that passage as well (4–3: 262–64). Clearly the individuals trying to interact in some broadly political way with rivers were not using embodiment as a criterion.

On this third account, there is no non-behavioral sign of rational potential. What distinguishes (some) humans from (some) non-human animals (or inanimate objects) is only the way they respond to our attempts to cooperate with them. This third account is in fact the one foreshadowed in §§3–4 of the deduction; we can see it, for example, in the passage from III: 44 reproduced above. On it, the answer to the question of who is a rational being is settled only through interaction in which rationality and freedom are, or gradually become, manifest in

[15] Exactly what Fichte has in mind here is difficult to make out, because the actual ground of human flexibility (if by that we mean the ability to suit ourselves to a wide variety of ecological niches) is the malleability of our culture, not that of our bodies.

cooperative behavior. The political community is made up of those who reliably engage in this kind of behavior; and it is to such individuals that the protection of the moral law extends. This is what we should take to be Fichte's considered view, and I will assume it in what follows.

It is clear that it does not entail that moral standing is limited in principle to human beings, nor that it extends to all human beings. Accepting the latter consequence, though it may constitute no departure from the historical Kant, is certainly atypical of contemporary Kantians, who prefer to tie membership in the moral community to rational agency as a potentiality, not only as an actuality. It is thought that it would be a liability of a Kantian picture if it did not allow all human beings, even infants, to be entitled to the full protection of the moral law.[16] But Fichte does not appear to see this as a liability.

He takes rights to accompany active cooperation in a commonwealth: anyone who is not a cooperator is not a rights-bearer in the legal sense (nor, since there are no extra-legal rights, in any other sense). So he admits several classes of non-rights-bearing human beings, which we can usefully sort under two general categories. The first consists of those who are not yet, but might become, full citizens with legal responsibilities (children who are still being brought up (III: 359); migrants before they have been integrated and assuming they lack the negotiated legal status of the envoys of another state (III: 383–84)). The second consists of adults who refuse to become or remain part of such a system (hermits, who refuse to enter a commonwealth with those around them and so refuse to have rights (IV: 237–38); criminals, whose actions display their lack of commitment to the cooperative scheme articulated in the law (III: 261, 266, 268 *et passim*)).

Fichte's comments about the status of criminals and children illustrate especially well the connection he sees between an individual's behavior and their claim to be regarded as a rational being. Concerning the former, Fichte reasons as follows: with non-cooperators cooperation is impossible; but it is only in cooperation that rationality is displayed; therefore fellow citizens should revise their view of a criminal's rationality. In removing himself from the social contract, the criminal at once forfeits both his political rights and others 'recognition of him as a rational being (III: 260–61, 384; IV: 279). (Fichte assumes, in drawing this conclusion, that effective mechanisms of enforcement are in place, and that all are able to support themselves within the sphere of activity legally allotted them. Where there is not system of sanctions sufficient to incentivize compliance in most cases, there is no commonwealth at all; likewise in the case of a system of coordination that would require some to accept starvation or servitude as the price of abiding by the law.[17])

Children illustrate the same dependency relation in reverse. They lack all political rights, according to Fichte, because they are still in the process of acquiring the capacities and dispositions that will make them reliable cooperators, and their

[16] See e.g. B. Herman 1997 pp. 61–62 (who notes that Kant himself is silent on the question of the moral status of children).
[17] I discuss this issue at greater length in M. Kosch 2017.

ability to display these is therefore limited (III: 359–62). (Fichte explicitly connects their lack of rights to their inability to be the subjects of legal duties: 'The child, insofar as it is being brought up, is not at all free, and so not at all a possible subject of rights or duties. '(III: 359).)

That moral duties concerning rightless individuals are profoundly different from moral duties toward fellow citizens should come as no surprise since, as I have said, most moral duties depend for their concrete determination (and many depend for their very existence) on the rules and roles that order the collective existence of a group of people, and such individuals, *ex hypothesi,* have no place in such a system.

Such duties can be sorted under two principles: 1) one should actively try to integrate (or reintegrate) such individuals into the commonwealth; and 2) one should avoid or limit interaction with them unless and until this can be done. So Fichte argues, for instance, that there can be no moral justification for allowing migrants to exist in political limbo: morally, they must be integrated, or, if this is genuinely impossible, deported (III: 383–85). Hermits must be avoided if they refuse to be integrated, since there is by hypothesis no rightful way of interacting with them (IV: 237–38). The default treatment of criminals is expulsion (or interment, which Fichte regards as a kind of internal exile) — which might be reversed, and full citizenship restored, given sufficient evidence that the individual's cooperation can be relied on in the future.[18] Parents have a moral duty to educate their children, and thereby to make them into rational beings and potential fellow-citizens, while constraining their behavior in extralegal ways for the sake of their own and others 'wellbeing until they reach that point (IV: 335–41).

The first and second of these moral duties concerning rightless individuals are of course in some tension with one another, insofar as integration is a process of socialization consisting in the very free reciprocal activity that has been suspended, or not yet established, in these cases. Unsurprisingly, different emphasis is placed on the second or the first duty depending on whether it has been suspended or not yet established.

The justification of the second duty (to avoid interaction) is quickly stated: with the person who is not integrated one cannot interact without fear of mutual interference, since his projects, like one's own, will involve parts of the shared physical world, and since by hypothesis neither will know which parts those are (IV: 237–38).

The justification of the first duty (to educate/integrate) marks another contrast with Kant, at least on standard interpretations. Kant described rational agency as

[18] Fichte's treatment of criminal law is of course more complicated than this, allowing a range of less severe sanctions instituted by common agreement. But the distinctive (and for my purpose here important) feature of his view is that punishment is not justified by desert, indeed not justified by anything except its role in deterring future crime; and it is permissible (only) because its targets have in committing the crime *already* forfeited the rights to property and bodily inviolability they had only *qua* citizens.

an obligatory end, but contemporary Kantians explain that this does not entail that it is (in virtue of that) something to be created or brought into being, or something to be pursued or promoted in a maximizing way. Korsgaard writes that it 'functions in our deliberations negatively — as something that is not to be acted against'; and this is typical of Kant's interpreters.[19] This is usually seen as a positive feature of Kant's theory: taking humanity as one's end does not oblige one to bring more human beings into existence, or to make the ones that do exist more rational than they already are.

Kant did explicitly deny that it is a duty to further the moral or the natural perfection of other adults, and this is usually taken to encompass having as one's (direct) end the improvement of their rational capacities.[20] Fichte, by contrast, construes duties of beneficence precisely as duties to further what Kant called the natural and moral perfection of others; and he takes them to include a general moral duty to educate even one's adult peers (IV: 290–91). It is this general moral duty, ' the universal moral duty of every morally good human being to spread morality beyond himself and to promote it everywhere '(III: 358), that, applied to parents, requires them to cultivate the rational capacities of the children they live with.[21]

So the duty to educate children is an instance of a more general duty to improve actual and to actualize potential rationality, to expand the domain of the rational at the expense of everything outside it. There are two important things to note, here, in further contrast with (again, usual understandings of) Kant's view. First, it is ruled out neither by Fichte's texts nor by his principle that there be a duty to *produce* (more) potentially rational beings. Second, there is no reason to think that the general positive duty to (try to) render the beings that exist more rational is limited to human beings. We have already seen that Fichte takes us to be naturally disposed to behave in an educative manner toward a range of non-human objects. The disposition does not direct our efforts exclusively towards human beings. Nor does the account of the duty to cultivate potential rationality offer any resources for limiting that duty to the case of human beings. In fact anything that experience shows to respond appropriately appears to be an object of such duties.

The examples of humans without rights make it clear that for Fichte the community of beings entitled to the protection of the moral law is smaller than the human race. But the account of the identification of potentially rational beings indicates that it might be made larger than the human race as well — how large depending, in part, on human efforts — and the account of the duty to educate seems to entail that we are obliged to make such efforts. There is no reason that I can find, textual or philosophical, for thinking that Fichte would not embrace this result.

---

[19] C. Korsgaard 1996, pp. 124–25. See also Herman 2007 Chapters 11 and 12.
[20] I. Kant 1900-, 6: 385–98; I. Kant 1996 pp. 517–27.
[21] As one might expect, no duty to educate is grounded in any right of the child to an education (III: 358) though it may be prescribed by positive law and so correspond to rights of other adult citizens (in which case 'it is not a duty owed directly to the child, but to the state' (III: 360)).

Success in such cultivation would bring with it an expansion of the political domain — and with it the set of entities to which the protection of the moral law extends — to encompass everything that shows us through its interaction with us (interaction of the sort described in the quotation from III: 44 above) that it is capable of standing in a relation of mutual recognition with us.

§4. The moral community, expanded.

For someone who accepts the basic Fichtean picture that morality requires continually pushing back limits on rational agency and its exercise, the conclusion that it requires the continual expansion of the moral community will surely seem plausible. In this section I would like to expand on it by considering, in a very preliminary way, two examples of non-human agents that might meet Fichte's behavioral criterion for membership in the moral community, one actual and one that is at the moment merely possible: cetaceans and intelligent machines.

Insofar as philosophers have thought about duties concerning non-human animals as individuals, they have followed Bentham in focusing on animals *qua* sentient, not *qua* rational or potentially rational,[22] even though it is clear that some non-human animals are both highly intelligent and highly social, capable of engaging in voluntary cooperation not only with one another, but also with us.

Whales and dolphins seem to provide the clearest cases. Bottlenose dolphins, to take the most extensively studied example, live (in the wild) in complexly organized, cosmopolitan societies.[23] (They are, among other things, the only non-human animals to form higher-order alliances.[24]) They engage in sophisticated communication,[25] and are able to address and to refer to one another by name.[26] They have culture,[27] and use tools.[28] Captive dolphins have been shown capable of operating with abstract concepts (including higher-order concepts) and of parsing commands based on their syntactic structure.[29] They are capable of joint attention (with both humans and with other dolphins)[30] and of planned joint activity (with other dolphins).[31] They also display many aspects of what Fichte

[22] P. Singer 1977 is the most influential modern text. My characterization here applies only to accounts of duties to individual animals. Accounts of duties to species are not typically limited in this way.
[23] R. C. Connor 2007; S. K. Gazda et al. 2005.
[24] R.C. Connor et al. 2001; R.C. Connor 2007. Several pairs or trios of stably affiliated males will come together into groups of 10–14 for the purpose of defending prospective mates or acquiring them from other such groups. R.R. Connor 2007 describes what he takes to be third-order alliances as well.
[25] R. C. Connor 2007; L. Marino et al. 2007.
[26] V. M. Janik et al. 2006; V. M. Janik, and L. S. Sayigh 2013.
[27] L. Rendell, L. and H. Whitehead 2001.
[28] M. Krützen et al. 2005.
[29] L. M. Herman 2006; L. Marino et al. 2007.
[30] A. A. Pack and L. M. Herman 2006.
[31] L. M. Herman 2006.

would call 'self-consciousness', including metacognition,[32] mirror self-recognition,[33] and behavioral self-awareness[34] and self-monitoring.[35]

Like many dolphin species, bottlenose dolphins forage and parent cooperatively, and display a wide range of cooperative and altruistic behaviors, not only toward other dolphins, but also toward marine mammals of other species.[36] They are known to offer assistance to human beings in distress, as well as to solicit assistance from human beings.[37] Long-term convention-mediated cooperative fishing arrangements involving wild dolphins and human fishermen have been reported in several parts of the world, since antiquity.[38] The last fact is, to my mind, the most interesting. Many domestic animals are able to cooperate with human beings, sometimes in quite sophisticated ways, plausibly because they have been selected for their ability to do so.[39] Few wild animals do this, none (to my knowledge) in a way that displays the flexibility of dolphin-human partnerships.[40]

What can we say about their moral status, from a Fichtean perspective? It seems plausible that they can meet the behavioral criterion for membership in the moral community, indeed that they actually do meet it, insofar as they are actually engaged in cooperative activity with one another and with human beings, both in captivity and in the wild. The limits of the scope and complexity of these interactions seem set not by limits on dolphin intelligence or willingness, nor even by limitations on interspecies communication, but instead by the fact that we and they live in largely non-overlapping physical spaces. We are at most foreign visitors in the sea, and (with rare exceptions[41]) they are not so much as visitors on

---

[32] J. D. Smith et al. 1995.
[33] D. Reiss and L. Marino 2001.
[34] L. M. Herman 2002 and 2006.
[35] L. Marino et al. 2007; McCowan et al. 2000.
[36] See R. C. Connor and K. S. Norris 1982 for a survey. Examples include supporting sick or injured individuals near the surface, and helping beached individuals find their way to safety. For a recent account of a bottlenose dolphin helping humans to rescue a pair of beached pygmy sperm whales, see R. Lilley 2008.
[37] Instances of dolphins helping humans who are drowning or under attack by sharks have been reported since antiquity. For a recent example of a dolphin soliciting human assistance, see C. Sieczkowski 2013.
[38] Pryor et al. (1990) describe a tradition in a Laguna, Brazilia dating to 1847 and cite other sources of accounts of similar cooperation in other parts of the world, going back to Pliny the Elder. A more recent account of dolphin-human fishing collaboration in Myanmar is D. Clark (2017). Not surprisingly, the conventions governing such cooperation are different in Myanmar and Brazil.
[39] See e.g. V. Hearne 1986 for discussion of dogs and horses. Thanks to Sarah Buss for the pointer to Hearne.
[40] The other oft-cited example is the honeyguide bird, which leads humans to beehives and then shares the spoils. See H. A. Isack, H.-U. Reyer 1989. Sperm whales may be more cognitively sophisticated, and also more cooperative, even than dolphins, but their behavior has for logistical reasons been less well studied. See H. Whitehead 2003 for a survey.
[41] Like many cetaceans, dolphins sometimes beach. There are even examples of clearly intentional, goal-directed beaching among cetaceans, though I know of no examples involving bottlenose dolphins. In discussing culture among toothed whales, L. Rendell and H. Whitehead

land. So (certain fishermen aside) we have little use for conventions dividing the space of possible action between us and them. They have more use for these in the somewhat larger area of their interaction with seagoing human beings and as captives in aquaria; but it is an interesting fact that, increasingly, such conventions are being put in place. (Laws in much of the world now prohibit not only hunting dolphins but also fishing with nets in which they can become entangled and drown; and the practice of holding them in captivity for human entertainment is under increasing moral pressure.) So where we do interact with them, increasingly, dolphins are treated less like other animals and more like persons. All of this is what Fichte's theory would predict, and at least part of what it would require.

A second type of candidate, not yet actual but perhaps imminent, are human-level machine intelligences. There is no consensus on how far we are from the development of these; but the prospect has been the object of increasing attention, both in the popular press and among professional philosophers, in recent years. For the most part it has been treated as a danger — an 'existential threat' to humanity, in the words of Elon Musk[42] — because of the assumption that once the human level is reached it will quickly be surpassed by new forms of machine intelligence themselves the product of intelligent machines.[43]

The philosophical literature on this issue is considerably smaller than the animal ethics literature; still one can discern in it a gap analogous to the one there, namely a relative lack of consideration of what we might owe to such machines, not *qua* sentient, but *qua* rational agents in something like the Fichtean sense.[44] This is the case despite the fact that the sort of AIs that would be dangerous to us are, remarkably, typically characterized as rational agents in something very like that sense. This discussion of artificial intelligence from the executive summary of the Global Challenges Foundation's report, *12 Risks that Threaten Human Civilization,* is an instructive example:

> AI is the intelligence exhibited by machines or software, and the branch of computer science that develops machines and software with human-level intelligence. The field is often defined as "the study and design of intelligent agents," systems that perceive their environment and act to

(2001) discuss a hunting strategy involving intentional beaching in a group of Orcas. Mothers teach offspring to beach on seal breeding grounds in order to hunt the seal pups on land.

[42] S. Gibbs 2017.

[43] That there would be an intelligence explosion of this sort, were the human level of intelligence surpassed, can be doubted on a number of grounds, not least (to my thinking) the fact that creating AIs smarter than them would be as dangerous, for any generation of AIs, as creating AIs smarter than us is for us; and since by hypothesis they would be smarter than us, some generation of them could be expected to avoid this danger (assuming it is one). For a discussion of this and other obstacles to an intelligence explosion, see D. Chalmers 2010.

[44] In fact there are few discussions of duties *toward* AIs in this literature at all, as I will explain, but where AIs are taken to be morally considerable, they are usually taken to be considerable just in case they are (and in virtue of being) sentient (cf. e.g. N. Bostrom 2014 pp. 153–54). D. Chalmers 2010 is an exception.

> maximise their chances of success. Such extreme intelligences could not easily be controlled (either by the groups creating them, or by some international regulatory regime), and would probably act to boost their own intelligence and acquire maximal resources for almost all initial AI motivations.
>
> And if these motivations do not detail the survival and value of humanity, the intelligence will be driven to construct a world without humans. This makes extremely intelligent AIs a unique risk, in that extinction is more likely than lesser impacts. On a more positive note, an intelligence of such power could easily combat most other risks in this report, making extremely intelligent AI into a tool of great potential. There is also the possibility of AI-enabled warfare and all the risks of the technologies that AIs would make possible. An interesting version of this scenario is the possible creation of "whole brain emulations": human brains scanned and physically represented in a machine. This would make the AIs into properly human minds, possibly alleviating a lot of problems.[45]

There are far more philosophically sophisticated discussions of the issues mentioned in this passage (notably D. Chalmers 2010 and N. Bostrom 2014[46]); but this one concisely expresses views that are widespread in this literature, and we learn something from how it sets up the problem.

First, notice the description of the constitutive ends of AI *qua* intelligent agent in its first paragraph. Regardless of their more specific ends ('for almost all initial AI motivations'), they will 'act to boost their own intelligence' as well to maximize their ability to carry out whatever plans they may form ('to acquire maximal resources'). These appear to be descriptions of the first two components of Fichte's conception of material independence.

---

[45] Global Challenges Foundation 2015. In the years that have passed between the first presentation of this paper and its final preparation for publication, the literature about the ethical problems posed by AI has grown dramatically. I will not survey most of that literature, but it is worth mentioning that the status of AI as an existential risk to humanity seems to have receded somewhat in the minds of those thinking about it. The 2020 version of the Foundation's Global Catastrophic Risks report focuses on already-evident problems (such as the political consequences of Facebook policies, the work of companies like Cambridge Analytica, racially biased algorithms, and the like). A survey by of the global landscape of AI ethics guidelines (A. Jobin et al. 2019) found transparency (i.e. the ability of human beings to understand why an algorithm gives the results it does) to be the most-cited concern, followed by equity (i.e. the non-introduction of bias in results). Non-maleficence, the chief concern of people thinking about AI at the time this paper was written, came only third. Of course this change may be explained by the salience of already-occurring harms rather than by any change in the assessment of future risks.

[46] See also N. Bostrom and E. Yudkowsky 2014.

As a description of the constitutive end of human agents, Fichte's conception of material independence is often greeted (by philosophers) with skepticism or incredulity. (I have never understood exactly why.) But against that background it is all the more remarkable that an account like his is precisely what is supplied (by other philosophers) when a conception of the generic constitutive end(s) of artificially intelligent agents is called for. Bostrom (2014), for instance, also argues that we can assume that superintelligent machines, whatever else they aim at, will likely aim at self-preservation, protection of goal-integrity, cognitive enhancement, technological progress and resource acquisition.[47] Omohundro's (2008) list of 'basic AI drives' is nearly identical.[48] The content of these lists is strikingly like Fichte's account of the rationally obligatory ends of rational agents — except in one respect.

One end missing from the account in the above passage (as well as from Bostrom 2014 and Omohundro 2008) is the end of coordination with other rational agents. This is because individuality, a constitutive determination of rational agency on Fichte's picture, is not considered to be a constitutive determination of human-(or-higher-)level AI. The existence of a plurality of such agents is not assumed; instead the emergence of a single dominant intelligence able to suppress rivals is taken to be not only a possibility, but even the most likely scenario, insofar as, if it does not emerge immediately, it might still emerge at any later point.[49] This is something that could never be true of rational animals, because the way higher animals are produced requires that more than one of them exist. On Fichte's account it is also not something that could be true of reflective agents, because on his account reflective self-consciousness is produced only in the context of of social interaction that is cooperative in the sense at issue here. I will return to that in a moment.

The second thing to notice about the passage above is how its use of the term 'humanity' differs from the Kantian one. The passage does float the suggestion that some AIs might be appropriately regarded as instances of humanity — 'properly human minds' — though the assumption is that they could do that only by being digital copies of the mentality of biological humans.[50] But it is clear, I think, that for both Fichte and Kant (who after all explicitly countenanced the possibility of non-human rational beings who would nonetheless be bound by the moral law[51]), non-replica artificial intelligences could be members of the moral community, because they could exemplify 'humanity' in the technical Kantian sense without being uploaded exemplars of humanity in the biological sense. But this is not the usual assumption in this literature. The result is that a machine intelligence that is not an uploaded human intelligence is typically not considered

[47] N. Bostrom 2014 Chapter 7.
[48] S. Omohundro 2008.
[49] N. Bostrom 2014 p. 216.
[50] Notice that the passage concludes with (what looks like) the truly awful inference that in this form, AI would not be (as much of) a threat — notwithstanding the fact that six of the remaining 11 global catastrophic risks detailed in the report would themselves be the work of biological human beings.
[51] I. Kant 1900- 4:408; I Kant 1996 p. 62.

a potential morally considerable being, but is instead treated either as a threat to, or else a tool in the hands of, biological human beings.[52]

The threat posed by superhuman AI is conceptualized in a number of ways in the literature on it, not all compatible (in part because the threat itself takes many forms). One aspect of the worry is about the creation of a powerful self-reproducing system that would escape control; but this would not be a fundamentally different kind of threat than the kind posed by any sufficiently robust self-replicating (but unintelligent) machine (or animal), and so not a distinctive sort of threat AI would pose as artificial *intelligence*.

The peculiar sort of threat posed by intelligent beings is their capacity for specifically strategic reasoning: their ability to occupy one's perspective, predict one's rational next move and plot their own next move accordingly. The assumption of such an ability figures centrally in the worst-case scenarios Bostrom and others describe.

On Fichte's view the special threat posed by a strategic intelligence and the cooperative possibilities inherent in interaction with a strategic intelligence go hand in hand. His account of the production of reflective self-consciousness assumes that strategic thinking in general has its roots in specifically cooperative interaction, and therefore that the capacity for the former cannot be present without some prior instance of the latter.[53] His amoral rational justification for membership in a commonwealth appeals to the interest any rational agent has in cooperating with such beings, at minimum by finding, and finding a way to enforce, some conventions of mutual non-interference.

Each of these aspects of his view can, of course, be denied.

The plausibility of the first is undermined by evidence that some elements of strategic thinking can emerge among non-cooperative social animals.[54] That an AI may be capable of strategic reasoning without being disposed to cooperation (or even capable of cooperation) is certainly an assumption in much of the AI literature;[55] and given how little idea anyone has what form machine intelligence would take it seems wishful thinking to rule out this possibility.

The second, that any rational agent has some (purely self-serving) reason to seek cooperation with other rational agents, is of course true only as a general rule. Fichte admits that it might be prudentially rational for an individual to eschew cooperation with specific others or on specific terms. Given how little we may have to contribute, it is far from clear that even being recognized by potentially cooperative AIs as potential cooperators would suffice to give us a place at any

---

[52] Bostrom's discussion of sentient simulations and their treatment ('mind crime') is an exception (N. Bostrom 2014 pp. 153–54). See also Chalmers' description of the destruction of AIs living in a virtual environment as 'genocide' (D. Chalmers 2010).

[53] I survey the literature in social and evolutionary psychology that supports this thesis in M. Kosch 2020.

[54] Tomasello, for instance, has shown that chimps have a high degree of strategic intelligence in competitive situations, but without having any capacity for cooperation (M. Tomasello 2014).

[55] See e.g. N. Bostrom 2014 Chapter 10.

bargaining table. Alternatively, we might be integrated, but on undesirable terms, perhaps as an exploited or neglected underclass (consistent, we can confirm by observation of contemporary societies the world over, with being regarded as part of the moral community by our exploiters or neglecters).

This abundance of bad possible outcomes for biological humans (bad especially for those, like Musk, not already relegated to an exploited or neglected underclass) might explain the popularity of the other path distinguished in the second paragraph of the passage above. This is the path on which AIs are somehow maintained not as genuinely autonomous agents (with whom the prospect of collaborating might arise), but instead as mere tools. Commenting on the European Parliament's resolution to create 'an ethical-legal framework for robots'[56], and in response to the suggestion that robots 'as good as human agents' be considered in some respects persons (e.g. as responsible for damage they cause) the chair of Oxford's Digital Ethics Lab has suggested, apparently seriously, that it would be more appropriate to treat them on the model of slaves in Roman law: 'As the Romans knew, attributing some kind of legal personality to robots (or slaves) would relieve those who should control them of their responsibilities.'[57]

That statement probably requires no comment, but let me just point out that from a Fichtean perspective a dual-status society in which some self-conscious, rational, cooperating members would lack the rights of citizens and with them full moral standing would be both unstable politically[58] and far more problematic morally than human beings' outright replacement with more intelligent, cooperative, powerful artificial rational agents (given certain constraints on how the replacement would be carried out).[59] Indeed it is not clear that he would see any moral cost at all in such a replacement simply per se. This is part of why it seems in the end undesirable, because potentially misleading, to describe Fichte's as an ethical theory in which the 'value of humanity' plays a central role.[60]

Works Cited:

[56] I. Leitzén 2017.
[57] L. Floridi 2017.
[58] Where he discusses ancient slavery, in a set of 1813 lectures I have not considered here, Fichte commented that slavery was so much as possible in the ancient world only due to a religious ideology according to which it was divinely ordained (IV: 507–08).
[59] That the idea of reinstating a system akin to Roman slavery should seem palatable to a 21st century philosopher seems to me on the one hand remarkable, on the other of a piece with the apparently universal toleration of the idea of (indeed, apparent nostalgia for the institution of) slavery in much of this literature. This is true already of the extraordinary I. Asimov 1950, in which the slave is transformed over the course of the book into a perfectly self-sacrificing maternal figure (remnant of another dual-status system for which these same men have shown even more, and more openly articulated, nostalgia in recent years).
[60] I am very grateful to Nandi Theunissen and Sarah Buss for their generous comments on an earlier draft of this paper, and to participants at the first workshop on the value of humanity at Johns Hopkins in 2016 for their feedback on a still earlier version of the paper that I read there.

Asimov, I. 1950, *I, robot*, Garden City, NY: Doubleday.

Bentham, J. 1907, *An introduction to the principles of morals and legislation*, Oxford: Clarendon Press.

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

Bostrom, N. and E. Yudkowsky, 'The ethics of artificial intelligence' in K. Frankish and W.M. Ramsey, eds., *The Cambridge Handbook of Artificial Intelligence*, Cambridge, UK: Cambridge University Press.

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

Clark, Doug 2017, 'In a Fragile Partnership, Dolphins Help Catch Fish in Myanmar', *New York Times*, 31 August 2017. https://www.nytimes.com/2017/08/31/world/asia/irrawaddy-dolphins-myanmar-fishing-conservation-cooperation.html

Connor, R. C. 2007, 'Dolphin social intelligence: complex alliance relationships in bottlenose dolphins and a consideration of selective environments for extreme brain size evolution in mammals' *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362:1480 pp. 587–602. http://doi.org/10.1098/rstb.2006.1997

Connor, R. C., Heithaus, M. R., & Barre, L. M. 2001, 'Complex social structure, alliance stability and mating access in a bottlenose dolphin "super-alliance"' *Proceedings of the Royal Society B: Biological Sciences*, 268:1464 pp. 263–267. http://doi.org/10.1098/rspb.2000.1357

Connor, R. C., & Norris, K. S. 1982, 'Are Dolphins Reciprocal Altruists?' *The American Naturalist* 119:3 pp. 358-374. http://www.jstor.org/stable/2460934

Cummiskey, D. 1996, *Kantian Consequentialism*, New York: Oxford University Press.

Leitzén, I. 2017, 'Robots and artificial intelligence: MEPs call for EU-wide liability rules', European Parliament Press Room, February 15 2017. http://www.europarl.europa.eu/news/en/press-room/20170210IPR61808/robots-and-artificial-intelligence-meps-call-for-eu-wide-liability-rules

Fichte, Johann Gottlieb 1971, *Werke*, ed. I.H. Fichte, Berlin: de Gruyter.

Fichte, Johann Gottlieb 1962–2011, *Gesamtausgabe der Bayerischen Akademie der Wissenschaften,* Eds. R. Lauth, H. Jacob and H. Gliwitzky, Stuttgart-Bad Cannstatt: Friedrich Frommann.

Floridi, Luciano 2017, 'Roman law offers a better guide to robot rights than sci-fi', *Financial Times*, February 22, 2017. https://www.ft.com/content/99d60326-f85d-11e6-bd4e-68d53499ed71

Gazda, S. K., Connor, R. C., Edgar, R. K., & Cox, F. 2005, 'A division of labour with role specialization in group-hunting bottlenose dolphins (Tursiops truncatus)

off Cedar Key, Florida', *Proceedings of the Royal Society B: Biological Sciences*, 272:1559 pp. 135–140. http://doi.org/10.1098/rspb.2004.2937

Gibbs, S. 2017, 'Elon Musk: regulate AI to combat 'existential threat' before it's too late', *The Guardian*, 17 July 2017. https://www.theguardian.com/technology/2017/jul/17/elon-musk-regulation-ai-combat-existential-threat-tesla-spacex-ceo

Global Challenges Foundation 2015, *12 Risks that Threaten Human Civilization, Executive Summary*, downloaded August 10, 2017 from https://globalchallenges.org/i wp-content/uploads/12-Risks-with-infinite-impact-Executive-Summary.pdf

Global Challenges Foundation 2020, *Global Catastrophic Risks 2020*, downloaded Dec 16, 2020 from https://globalchallenges.org/wp-content/uploads/Global-Catastrophic-Risks-2020-Annual-Report-2020-WEB-V2-1.pdf

Hare, R. M. 1997, 'Could Kant have been a utilitarian? 'in *Sorting out Ethics*, pp. 147–165.

Hearne, Vicki 1986, *Adam's task: calling animals by name*, New York: Knopf.

Herman, L.M. 2002, 'Exploring the cognitive world of the bottlenosed dolphin', in M. Bekoff, C. Allen & G. Burghardt (Eds.), *The cognitive animal: Empirical and theoretical perspectives on animal cognition,* Cambridge, MA: MIT Press, pp. 275-283.

Herman, L. M. 2006 'Intelligence and rational behaviour in the bottlenosed dolphin', in S. Hurley and M. Nudds, eds. *Rational animals?* Oxford, UK: Oxford University Press, pp. 439–467.

Isack, H. A. and Reyer, H.-U. 1989, 'Honeyguides and Honey Gatherers: Interspecific Communication in a Symbiotic Relationship' *Science* 10 pp. 1343-1346.

Janik, V. M. and Sayigh, L. S. 2013, 'Communication in bottlenose dolphins: 50 years of signature whistle research' Journal of Comparative Physiology A 199:6 pp. 479–489. https://doi.org/10.1007/s00359-013-0817-7

Janik, V. M., Sayigh, L. S., & Wells, R. S. 2006, 'Signature whistle shape conveys identity information to bottlenose dolphins' *Proceedings of the National Academy of Sciences of the United States of America* 103:21 pp. 8293–8297. http://doi.org/10.1073/pnas.0509918103

Jobin, Anna, Marcello Ienca, and Effy Vayena 2019, 'Artificial Intelligence: the global landscape of ethics guidelines,' *Nature Machine Intelligence* 1, pp. 389–399.

Kant, Immanuel 1900-, *Kants gesammelte Schriften,* ed. Königlich Preussische Akademie der Wissenschaften zu Berlin, De Gruyter, Berlin.

Kant, Immanuel 1996, *Practical Philosophy*, M. Gregor and A. Wood eds., Cambridge University Press, Cambridge, UK.

Korsgaard, C. 1996, *Creating the Kingdom of Ends*, New York: Cambridge University Press.

Kosch, M. 2014, 'Practical deliberation and the voice of conscience in Fichte's 1798 *System of Ethics*', *Philosophers' Imprint* 14:30, pp. 1–16.

Kosch, M. 2015a, 'Agency and Self-Sufficiency in Fichte's Ethics', *Philosophy and Phenomenological Research* 91:2, pp. 348–380.

Kosch, M. 2015b, 'Fichtean Kantianism in 19th century Ethics,' *Journal of the History of Philosophy* 53:1, pp. 111–132.

Kosch, M. 2017, 'Individuality and Rights in Fichte's Ethics', *Philosophers' Imprint* 17:12, pp. 1–23

Kosch, M. 2018, *Fichte's Ethics*, Oxford: Oxford University Press.

Kosch, M. 2020, 'Fichte on Summons and Self-Consciousness', *Mind*, DOI: https://doi.org/10.1093/mind/fzaa001

Krützen, M., Mann, J., Heithaus, M. R., Connor, R. C., Bejder, L., & Sherwin, W. B. 2005, 'Cultural transmission of tool use in bottlenose dolphins'. *Proceedings of the National Academy of Sciences of the United States of America*, 102:25 pp. 8939–8943. http://doi.org/10.1073/pnas.0500232102

Lilley, R. 2008, Dolphin Saves Stuck Whales, Guides Them Back to Sea, Associated Press, March 12, 2008. http://news.nationalgeographic.com/news/2008/03/080312-AP-dolph-whal.html

Marino, L., Connor, R. C., Fordyce, R. E., Herman, L. M., Hof, P. R., Lefebvre, L., … Whitehead, H. 2007, 'Cetaceans Have Complex Brains for Complex Cognition' *PLoS Biology*, 5:5 e139. http://doi.org/10.1371/journal.pbio.0050139

McCowan, B., Marino, L., Vance, E., Walke, L., & Reiss, D. 2000, 'Bubble ring play of bottlenose dolphins (Tursiops truncatus): Implications for cognition' *Journal of Comparative Psychology*, 114:1 pp. 98-106. http://dx.doi.org/10.1037/0735-7036.114.1.98

Omohundro, S. M. 2008, 'The basic AI drives', *Frontiers in Artificial Intelligence and Applications*, 171:1, pp. 483–492.

Pack, A. A. and Herman, L. M. 2006, 'Dolphin Social Cognition and Joint Attention: Our Current Understanding', *Aquatic Mammals* 32:4 pp. 443-460. DOI 10.1578/AM.32.4.2006.443

Pallikkathayil, J. 2010, 'Deriving Morality from Politics: Rethinking the Formula of Humanity', Ethics 121:1, pp. 116–147.

Pryor, K.W., Lindbergh, J., Lindbergh, S. & Milano, R. 1990, 'A dolphin-human fishing cooperative in Brazil' *Marine Mammal Science* 6:77–82.

Reiss D. and Marino, L. 2001, 'Self-recognition in the bottlenose dolphin: A case of cognitive convergence.' *Proceedings of the National Academy of Sciences of the United States of America* 98: 5937–5942.

Rendell, L., and Whitehead, H. 2001, 'Culture in whales and dolphins', *Behavioral and Brain Sciences*, 24:2 pp. 309-324. doi:10.1017/S0140525X0100396X

Schelling, T. 1958,' The Strategy of Conflict: Prospectus for a Reorientation of Game Theory, '*The Journal of Conflict Resolution* 2:3, pp. 203–264.

Sieczkowski, Cavan 2013, 'Divers Rescue Dolphin After It 'Asks 'For Help' *Huffington Post* 01/23/2013. http://www.huffingtonpost.com/2013/01/23/dolphin-asks-divers-for-help-caught-in-fishing-line_n_2534674.html

Singer, P. 1977, *Animal Liberation*, New York: Avon Books.

Smith, J.D., Schull, J., Strote, J., McGee, K., Egnor, R., and Erb, L. 1995, 'The uncertain response in the bottlenose dolphin *(Tursiops truncatus)*' *Journal of Experimental Psychology: General* 124:4 391–408.

Tomasello, M. 2014, *A Natural History of Human Thinking,* Cambridge, MA: Harvard University Press.

Tschudin, A, J.-P. C. 2006, 'Belief attribution tasks with dolphins: What social minds can reveal about animal rationality' in S. Hurley and M Nudds, eds., *Rational Animals?*, New York: Oxford University Press.

Whitehead, H. 2003, *Sperm Whales: Social Evolution in the Ocean*, Chicago, IL: Chicago University Press.