Q3: Responsibility
v.0.0.2

Q3: Responsibility 1
- Test cases 2
  - Discriminatory bank algorithm 2
  - Gatekeeper algorithms 3
  - AI bad things ??
  - Self-driving car crashes ??
  - Boeing 737-max and MCAS development ??
- Intro 3
- Two questions 5
- When is someone morally responsible? 6
  - (1) Causal connection: Actus reus 7
    - Omissions 7
  - (2) Knowledge: Mens rea 7
  - Problems: Causal connection 9
    - Problem of many-hands 9
      - Therac-25 treatment machine ??
    - Temporal and physical distance 10
    - Development practices 11
      - Libraries 11
      - Methodologies 12
  - Problems: Mental states 12
    - Bugs 12
    - Biases 13
  - Problems: freedom ??
- Implications of moral responsibility? 14
  - (1) Reactive attitudes and evaluation are appropriate 15
    - Reactive attitudes 15
  - (2) Compensation / retribution / apology may be owed 16
  - (3) Punishment may be appropriate 16

Expanding the notion of responsibility ??
Forward and backwards looking conceptions ??
Responsibility, liability, accountability ??
Floridi and gatekeepers 16
Moral patients 17
Moral agents 17
Warm up 18
Starting points 19
No anthropocentrism 19
Observability 20
Terminology 20
Agents 22
(1) Interactivity 22
(2) Independence 23
(3) Adaptability 24
Moral agents 25
Are gatekeeper algorithms moral agents? 26
Moral agency without responsibility 28
Objections 31
Contra observables 32
Contra anthropocentrism 34
The asymmetry 34

1 Test cases

As usual, lets get some test cases on the table before we start.

1.1 Discriminatory bank algorithm

Human beings are fallible. It is very hard for even well-intentioned people to consistently apply a set of criteria in judging a wide range of cases. Machines are great at this. It's what they do.

Thus it seems like a great idea to take decisions which have historically been infected by biases like racism out of the hands of humans. The computer won't care if you are black,

white, asian, or latinx. It will only care whether you are statistically likely to pay back your mortgage.

However, underwriting requires judgment. We can't just apply a set of clearly defined standards. There are a lot of tradeoffs and decisions about risk that need to be made on the basis of lots of other information.

This is where machine learning techniques can be helpful. By providing the algorithm with a huge dataset about past mortgage decisions and defaults, it can extrapolate its own rules which mimic those of a competent underwriter.

If the problem isn't yet obvious, here it is: Historically, mortgage decisions have been distorted by racism. We won't get non-racist decisions by using that data to train the system.

Racism in; racism out.

1.2 Gatekeeper algorithms

Zeynep Tufekci writes that

> Algorithmic gatekeeping is the process by which such non-transparent algorithmic computational-tools dynamically filter, highlight, suppress, or otherwise play an editorial role—fully or partially—in determining: information flows through online platforms and similar media; human-resources processes (such as hiring and firing); flag potential terrorists; and more. {Anonymous:SbzjS1LM} pp.207-8

Let's call these gatekeeper algorithms.

2 Intro

Now that we have a grip on some of the harms that might be involved with the use of personal data, we can turn our third question

> (Q3) How should we assign blame when people are harmed by algorithmic uses of personal data?

If an identifiable individual caused the harm, this isn't terribly interesting or controversial.

The more difficult issues arise when we have corporate agents or actions undertaken by mixes of human and machine. Both the technology itself and the way that technology is created may enable novel difficulties in understanding moral responsibility. Doorn nicely summarizes some of the issues here and how they differ from traditional thinking about responsibility.

> The ethical literature…often assumes: (1) that it are individuals who act, (2) that the consequences of their actions are causally direct traceable, and (3) that these consequences are certain. None of these assumptions seem to apply to many of the ethical issues raised by modern technology and engineering. First, engineering and technology development typically take place in collective settings, in which a lot of different agents, apart from the engineers involved, eventually shape the technology developed and its social consequences. Second, engineering and technology development are complex processes, which are characterized by long causal chains between the actions of engineers and scientists and the eventual effects that raise ethical concern. Third, social consequences of technology are often hard to predict beforehand. " {Doorn: 2012ij} p.2

Still, it's hard to get a grip on what new problems might be arising for moral responsibility. Normally, when a corporation or other hierarchical organization misbehaves, we just push responsibly up to senior leadership or the board of directors. Why isn't that enough here?

Suppose our company is responsible for something bad. Pick your example. The plane crashed. Software was hacked. Data was lost.

From the victims' perspective, the company is responsible. They (try to) sue the company to redress their losses.

From the Board's perspective, this is straightforward. If it was bad enough, they fire the CEO. Job number one for their successor is demonstrating that the disaster will never happen again.

From the new CEO's perspective, this is straightforward: fire whichever Vice-President oversaw the department whence the mistake arose.

You're the new VP of that department. You want to demonstrate, both to the CEO and to your subordinates, that such errors are unacceptable. Whom do you fire? How do you spell out a policy which holds subordinates responsible for such errors?

If it's a single engineer's error — she neglected to check the size of a variable and opened the software to a buffer overflow[1] — that's easy. But what of the team that did the code review? Also easy. If they should've caught the error, fired. If some procedural quirk allowed the bug to sneak through, we tighten up our reviews.

But most of the time, it's not going to be this easy. Engineers solve engineering problems. If the problem wasn't stated in the specs given to them, how were they to know to solve it? What about problems that arise through the unforeseen interaction of very different components, created by different engineers? Obviously, we want engineers to be conscious of their job beyond what's in the specs of what they are working on; this is a key theme in both security and safety —that security and safety have to be everyone's job. But how far does this go? If an engineer points out a problem and it gets passed up the chain, isn't her job done?

On top of this is the reality that in many engineering disasters, each of the decisions which led to it was perfectly reasonable in isolation; it's only when they are combined within an engineering process that the danger arises. That starts to sound like the process itself is potentially responsible. But processes can't be morally responsible, right? Only people can be morally responsible, right? Right?

3 Two questions

When things go wrong we want to identify and blame those responsible. There's a flip side to this too. If you're an engineer working on a project, what are you morally responsible for? Are you only responsible for preventing harms which come from the particular component you are designing? Or are you responsible for ensuring the whole system doesn't cause harm?

Let's separate two questions:

(A) When is a person morally responsible for a harm?

(B) What does it mean for a person to be morally responsible?

---

1. This is a very dangerous vulnerability which allows the attacker to run code on the user's machine. Here's a good explainer video from computerphile: https://www.youtube.com/watch?v=1S0aBV-Waeo

The first question is about identification; the second is about implications. Think of an analogy with the structure of a criminal case. There are two questions: Did the defendant commit the crime? And, if they did, how should the be punished. That's what I have in mind here.

4 When is someone morally responsible?

The question of when someone is morally responsible for something is controversial.[2] However, most answers will claim that:

> (1) S is morally responsible for harm x only if S is causally connected to x

Obviously we don't want to blame people for things they didn't do. Thus there must be some sort of causal connection between the agent and the outcome for which she is morally responsible. However, usually we want to go further and require that S has some control over the outcome / events.

> (2) S is morally responsible for harm x only if S knew, should have known, or could have known that her actions may cause x

If the causal sequence that led to the outcome was completely and totally beyond a person's knowledge and anything they could've predicted, then it is unlikely that they are morally responsible. We can only hold people to what they can do, thus we cannot blame them for failing to do the impossible.

> (3) S is morally responsible for harm x only if S could have done otherwise.

This is the always intuitive and always difficult to pin down free choice requirement. If S had no ability to do otherwise, we generally would not hold her morally responsible. The exceptions are usually cases where she somehow culpably put herself in the position where she could not do otherwise. [C.f., duress]

All of these are tricky. Thus to give ourselves the most to work with, let's update (1) and (2) with some concepts from philosophy of law and jurisprudence. In particular, let's

2. See for example https://plato.stanford.edu/entries/computing-responsibility/

add the common components of a crime to the mix, since some of the extra tools we're going to need may have already been developed in those areas.

As we go, I will try to set out how these practices related to technology complicate things.

4.1 Causal connection: Actus reus

Let's understand the causal connection required in terms of the actus reus of a crime. Every crime requires you to do something —it involves either a wiggling of the body or a failure to wiggle.

Often we supplement this with the outcome of the wiggling. For example, the actus reus of murder is homicide: causing the death of another. If you do all that you needed to do to kill someone but they fail to die, it is not murder (though it is a different crime like attempted murder or aggravated battery)

4.1.1 Omissions

Notice that we included both overt actions and failures to act. The latter are often called omissions. In US criminal law, there are very few crimes where the actus reus is an omission —I've heard that there are only 2, one of which is failing to file your taxes.[3]

In tort law and in ethics more broadly there are culpable failures to act. Many think that you would've done something wrong if you fail to save someone's life when you could've done so at little or no cost to yourself. The possibility of being responsible via omissions is pretty important when we get to the sorts of harms caused by engineering.

4.2 Mental state: Mens rea

If you've watched a lot of police procedural TV shows, you may know that (almost) every crime requires you to have a corrupt mental state. [4] This is the mens rea of a crime.

---

3. 'Good Samaritan' laws which require people to aid others in need or face punishment are extremely rare and hard to apply. They could involve omissions as the actus reus, but might actually require some sort of overt act (e.g., driving away).
4. 'Almost' because there are some strict liability offenses where no mental state is required. This is more common with infractions (a different category from crimes), e.g., speeding tickets.

Thus instead of talking about whether a person knew that the harm would/could result, we can distinguish different kinds of culpable mental state. The distinctions often affect our judgments of the crime's severity.

Under the Model Penal Code there are basically 5 culpable mental states:

- Purpose: You are trying to do x
- Knowledge: You know that what you are doing is x (even if you aren't trying to do it)
- Recklessness: You are aware of a substantial risk of harm from doing x
- Negligence: A reasonable person would've known that x creates a substantial risk of harm.
- Wanton and Depraved Heart[5] (sometimes: Malignant and Abandoned Heart): Basically negligence, but what you did was so f-ed up that we want to punish you more severely.

Let's look quickly at how these different mental states affect culpability.

Homicide, killing another person, is the actus reus of several crimes. It is not on its own a crime. The crime of murder is committing a homicide with purpose or knowledge (or a wanton and depraved heart). If you know that what you are doing is killing someone or if you are trying to kill them, you are committing murder.

Manslaughter is reckless homicide. You aren't actually trying to kill the person, but you are doing something that you know has a substantial chance of killing them. There are a variety of gradations here. Many jurisdictions distinguish between voluntary and involuntary manslaughter; depending on the details it may be that voluntary manslaughter involves recklessness and involuntary manslaughter involves negligence.

Some jurisdictions have negligent homicide statutes. These get tricky. For example, (IIRC) New Jersey had a negligent homicide statute which tried to punish drug dealers when people they sold the drug to die. That was struck down as unconstitutional.

Finally, in many jurisdictions, if you do something so despicable that your negligence demonstrates an utter lack of concern for human life, you may be charged with murder. Someone who thinks it is a fun game to throw pieces of brick off a highway overpass

---

5. Obviously, this is the best phrase in law. Soooo much better than 'A frolic of one's own'

between the cars underneath may be charged with murder when someone dies. Their mental state was basically negligence —the whole point of the game was to not hit the cars and they thought they were so good at it that they wouldn't actually hit a car— but statutes make it murder.

With this more sophisticated understanding of what's baked into responsibility, let's see if technology actually poses new problems.

4.3 Problems: Causal connection

The causal connection required for responsibility gets complicated in several ways by technology.

4.3.1 Problem of many-hands

Suppose again that you are the new VP overseeing the division which created the problem. You want to credibly promise to the CEO that this problem will never happen again. That means figuring out who is responsible and taking appropriate corrective action.

There is an epistemic problem here. It's hard to know who made which choice, especially from the outside. This is the original provenance of the problem of many-hands.[6] To see the problem, consider this, from the Stanford Encyclopedia of Philosophy

> One classic example of the problem of many hands in computing is the case of the malfunctioning radiation treatment machine Therac-25....During a two-year period in the 1980s the machine massively overdosed six patients, contributing to the eventual death of three of them. These incidents were the result of the combination of a number of factors, including software errors, inadequate testing and quality assurance, exaggerated claims about the reliability, bad interface design, overconfidence in software design, and inadequate investigation or follow-up on accident reports. Nevertheless…it is hard to place the blame on a single person. The actions or negligence of all those involved might not have proven fatal were it not for the other contributing events. This is not to say that there is no moral responsibility in this case as many actors could have acted

6. The original term comes from
Thompson, D. F. (1980). Moral responsibility and public officials. American Political Science Review, 74, 905–916.

> differently, but it makes it difficult to retrospectively identify the appropriate person that can be called upon to answer and make amends for the outcome.[7]

Subsequent literature has focused on sorting out whether this is just an insider-outsider epistemic problem, or whether there is a more metaphysical problem as well.[8] I'm going to suggest that there is.

It very well may be that every individual decision was perfectly reasonable on its own. It may be that the responsible entity is not one individual but the entire team. Moreover, the team's responsibility may only be understandable when we zoom out to the organizational structure in which it operates. The policy guidance, engineering requirements, and other institutional factors may be crucial for understanding what is responsible.

To be sure, this is very strange. We are not wired to ascribe responsibility to abstract things —teams within organizational context— where we cannot reduce the blameworthiness to the individuals comprising the group. Indeed, much of the ethical literature on responsibility in this context has focused on expanding our notions of responsibility to improve engineering practice. If we focus on ascribing blame — backwards looking responsibility— in the way I've discussed, we will not prevent future problems.

Another response to this claim is to push the metaphysics further. Perhaps we need to expand our understanding of what can count as a moral agent. As so often when we are suspicious that the moral issues raised by technology somehow require an extravagantly revisionist metaphysics, Luciano Floridi is here to help. We'll get to him in a bit.

4.3.2 Temporal and physical distance

---

7. https://plato.stanford.edu/entries/computing-responsibility/
Associated references:
(Leveson and Turner 1993; Leveson 1995) (Nissenbaum 1994; Gotterbarn 2001; Coeckelbergh 2012; Floridi 2013),

8. See: van de Poel, I., Nihlén Fahlquist, J., Doorn, N., Zwart, S., & Royakkers, L. (2011). The Problem of Many Hands: Climate Change as an Example, *18*(1), 49–67. http://doi.org/10.1007/s11948-011-9276-0

Causation is also complicated by the temporal and physical distance between actions and events enabled by computing. Years may pass between the creation of the code and the result of the harm. This poses significant epistemic difficulties. Indeed, many of the things human moral psychology[9] has trouble dealing with —distance, bad things happening to faceless strangers, uncertainty— will be features of technology-enabled harms.

More importantly, the engineer often has no idea who the consumer of a product will be or how it may be used. But many products can be used in completely unintended ways. If the creator of the code/machine had no way of anticipating that some feature they included would be used in an unanticipated way which results in a harm, how are we to say that they are responsible?

4.3.3 Development practices

The ways technology is developed pose several challenges for responsibility.

4.3.3.1 Libraries

Software (and often hardware) development involves using libraries —code which handles a specific set of tasks— that others have developed. This is often a good thing. It both saves time and can promote security. Unlike when I first started programming, if you are developing a web app today it would be irresponsible to write it all from scratch. Web developers use frameworks that others have created.[10] Because lots of people use and test the framework, its bugs and security flaws can be rapidly detected.

For a small example of how heavy this reliance can be, in my grading app, the code I've written comprises about 10 MB. The PHP libraries comprise 175 MB and the JavaScript libraries comprise 338 MB.

However, this has the downside that developers are essentially stringing together black-boxes —they don't really know how their software is working. It is also problematic because you will only get the latest security updates if you update the libraries. But new

---

9. This shouldn't be a surprise. We evolved in relatively small groups in relative isolation. We are really good at moral evaluation of effects on people we know and can see; bad at those we can't. Obviously, that doesn't mean we can't find ways to overcome them.

10. For example, in web development, I use Laravel [ref] for all the server-side operations and Vue to help build the part that people see.

versions of libraries come with all sorts of changes, some of which may break your system. Take it from me, the misery of trying to figure out how a necessary update has created a weird bug is quite exquisite.

This means that the people creating a program are not entirely creating the program. More importantly, they often are not in a position to understand how the program works. This is a completely standard practice, and on the whole, a huge benefit.[11] But it does pressure both the causal connection and the knowledge required for responsibility.

4.3.3.2 Methodologies

Software teams have a lot of different strategies for building products. Methodologies like scrum and agile are very common. These involve breaking a system down into bite sized chunks so that each developer works on a series of small, often disconnected parts. The software can thus be continuously improved by progressively updating the chunks, allowing you to get to market faster.

In a mature organization, there will be a code review by other members of the team or managers to check over work before it goes into production. Though this is not always the case. Presumably, when there is a review process, the reviewers acquire some amount of responsibility. Similarly, an organization's decision not to have a review process may be important to assessing responsibility for harms.

4.4 Problems: Mental states

If responsibility requires awareness of what you're doing[12], the interaction of technology and people complicates responsibility.

For one, end users of technology often have wildly mistaken ideas about how a system works. Modern technology is like magic. Most of us are not magicians. Let's discuss a few ways this complicates things.

4.4.1 Bugs

Every system has bugs. A good development and testing regime will ensure that obvious problems created by bugs don't arise for users. Though many companies skimp on testing in order to get to market quickly. More importantly, many bugs will involve

---

11. It also allows people trained in philosophy to do things that 20 years ago would require significant training in math or computer science.
12. Or what a reasonable person would've been aware of, as in negligence.

completely unanticipated corner cases.[13] In some cases, those situations should've been anticipated. In others, there's really no way they could've been.

[ToDo: Examples]

Oftentimes, bugs which have consequences for end users will be completely hidden from users and developers by interfaces.

Obviously, if there's no practical way someone could've (or should've) known that their action will create harm, the mens rea required for responsibility won't be present; they can't be blamed.

4.4.2 Biases

Another important issue for the required mens rea is that people are very often wrong about how well machines function. We often either overweight or underweight their accuracy; we may believe the machine readout over our own eyes; or fail to do so when we should believe the machine.

For some examples,

> The opacity of many computer systems can get in the way of assessing the validity and relevance of the information and can prevent a user from making appropriate decisions. People have a tendency to either rely too much or not enough on the accuracy automated systems (Cummings 2004; Parasuraman & Riley 1997). A person's ability to act responsibly, for example, can suffer when she distrust the automation as result of a high rate of false alarms. In the Therac 25 case, one of the machine's operators testified that she had become used to the many cryptic error messages the machine gave and most did not involve patient safety (Leveson and Turner 1993, p.24). She tended ignore them and therefore failed to notice when the machine was set to overdose a patient. Too much reliance on automated systems can have equally disastrous consequences. In

13. From wikipedia: "a corner case (or pathological case) involves a problem or situation that occurs only outside of normal operating parameters —specifically one that manifests itself when multiple environmental variables or conditions are simultaneously at extreme levels, even though each parameter is within the specified range for that parameter."
https://en.wikipedia.org/wiki/Corner_case

> 1988 the missile cruiser U.S.S. Vincennes shot down an Iranian civilian jet airliner, killing all 290 passengers onboard, after it mistakenly identified the airliner as an attacking military aircraft (Gray 1997). The cruiser was equipped with an Aegis defensive system that could automatically track and target incoming missiles and enemy aircrafts. Analyses of the events leading up to incident showed that overconfidence in the abilities of the Aegis system prevented others from intervening when they could have. Two other warships nearby had correctly identified the aircraft as civilian. Yet, they did not dispute the Vincennes' identification of the aircraft as a military aircraft. In a later explanation Lt. Richard Thomas of one of the nearby ships stated, "We called her Robocruiser… she always seemed to have a picture… She always seemed to be telling everybody to get on or off the link as though her picture was better" (as quoted in Gray 1997, p. 34). The captains of both ships thought that the sophisticated Aegis system provided the crew of Vincennes with information they did not have.[14]

5 Implications of moral responsibility?

We've talked about the conditions of moral responsibility —when one is responsible-- and the challenges technology imposes. Let's turn now to what we mean by moral responsibility. If someone is morally responsible for a harm, what does that entail for how we may relate to her? May we call her mean names? Shun her? What does it entail for how she should relate to herself? Should she feel guilty? Should she be mad at herself?

Responsibility is a common notion which we use all the time. Thus we shouldn't be surprised that it is hard to pin down when we look closely at it; indeed, there are several related concepts which we normally lump together.

In this sort of situation, it's often helpful to get started by inverting the question. For example, here we ask what it means to say that someone's not responsible. Pretty clearly, it means they owe no apology, it would be wrong to punish them, and attitudes like blame would be inappropriate.

Thus we can say that moral responsibility means or implies at least three things

1) Reactive attitudes and evaluation are appropriate
2) Compensation / retribution / apology may be owed

14. https://plato.stanford.edu/entries/computing-responsibility/

3) Punishment may be appropriate

Some writers think that the we need to break the connection between responsibility and blameworthiness. That is a pretty radical suggestion which we'll get to below. First, let's take each of the three components

5.1 Reactive attitudes and evaluation are appropriate

Reactive attitudes are mental states which we take towards someone's actions. These include states like praising, blaming, criticizing, punishing, being disappointed, being pleased, et cetera. They matter for ethical evaluation because the appropriateness of experiencing them is governed by our moral norms. It is appropriate to feel grateful or direct praise toward someone who rescues children from a burning building. There is something wrong with a person that fails to feel angry or critical towards someone who gratuitously injures a child.

Thus when I say that moral responsibility involves reactive attitudes and evaluation being appropriate, I mean

> (1) S is morally responsible for x only if it is appropriate for S to be the subject of reactive attitudes like blame for x

If Scarlet saves a bunch of children from a burning building while Violet just watches, it would be appropriate to praise Scarlet. It would not be appropriate to praise Violet.[15] That's because Scarlet is morally responsible for the children-saving. Violet is not.

This is a necessary condition but not a sufficient condition because there may be other non-moral reasons for punishing /rewarding —e.g., rewarding the salesperson who sold the most last month. Though even in these cases, there may sill be some connection with moral responsibility because our sense of justice is tied to responsibility. If you actually sold more product but someone else gets the salesperson of the month award, you can legitimately complain that it is unfair. That's a moral complaint.

5.1.1 Reactive attitudes

15. NB, that doesn't necessarily mean that we should criticize Violet. Just the when we are making a list of heroes, Scarlet belongs on it, Violet doesn't.

The reactive attitudes which are appropriate can get tricky. Generally speaking, a person deserves blame when they've done something wrong.

[ToDo: issues with blame]

[ToDo: other reactive attitudes]

However, there are also cases of blameless wrongdoing. In these cases, a person may owe an apology or otherwise be responsible for repairing what has been broken.

5.2 Compensation / retribution / apology may be owed

If someone is not responsible for a harm, it would not make sense (or even be unjust) to demand that they compensate the victim. However, these forms of repair are appropriate when someone is responsible for the harm.

> (2) S owes compensation / retribution / apology to V for harm x only if S is morally responsible for x.

5.3 Punishment may be appropriate

It is unjust to punish someone who is not responsible for a wrong. Punishment here covers both legal punishment at the level of society —e.g., criminal sentences or punitive damages—and non-legal interpersonal forms —e.g., shunning.

> (3) S may be punished for x only if S is morally responsible for x.

Note that this is a place where judgments of moral responsibility diverge from causal claims. It is a necessary condition because it is possible to cause a bad thing without being morally responsible. This is why children are not treated the same as adults. It's also why justice requires the existence of an insanity defense.[16] If someone lacks the capacities for moral responsibility, it is wrong to punish them.

---

16. The insanity defense (as well as the defense of infancy —being a child) is different from other defenses in that it asserts that the defendant is not the sort of being which can be subject to the law.

6 Floridi and gatekeepers

Let's turn now to a fairly extravagant claim: that some computer programs may need to be regarded as moral agents responsible for harms.

6.1 Moral patients

Before getting into Floridi's picture, let's start with the notions of moral agents and patients. These concern what sort of beings 'count' in moral considerations.

Suppose that, deep in the wilderness, you smash a rock with a hammer. Since the rock doesn't belong to anyone, you do nothing wrong. Indeed, it seems weird to even ask about whether it was right or wrong. Whereas, if you then hit me with a hammer, the question of right or wrong is very natural. To capture what is at stake, we will say that

> x is a moral patient if and only if x's interests must be considered in moral decision-making

We do something wrong if we ignore the interests of a moral patient.

Rocks are not moral patients. We do not need to consider their interests in deciding whether to throw them.

Animals are moral patients. We need to consider their interests before we do things to them. Throwing a rock at a dog is wrong because of what it may do to the dog, not the rock.

6.2 Moral agents

We can ask about what sort of beings can act in moral/immoral ways. Occasionally, a primate study will make breathless headlines with claims like 'Chimps can act morally' . TV shows ask us to consider whether psychopaths are capable of morality. Mental illness and the insanity defense are common subjects of controversy. All of these are questions about when something counts as a moral agent.

Thus we will say that

> x is a moral agent if and only if x's actions can be subject to moral assessment

By moral assessment, I mean that it is appropriate to use moral concepts in judgment of x's actions. Normally, I would put this differently: moral agents can be morally

responsible. However, as we'll see below, Floridi wants to say that moral agents may not be morally responsible. Thus we need the wider definition.

Volcanos are not moral agents. When a volcanic bomb —a rock blasted into the air— hits and injures a person, it makes no sense to ask if the volcano did something wrong. It's not the sort of thing which can do wrong.[17]

Similarly, most animals are not moral agents. Some species may be, the jury is still out. Often, when we seem to apply moral concepts to animal actions —we say that the cat playing with the doomed mouse is cruel— we don't really mean it, as is shown by our immediately inviting the 'cruel' beast into our home.[18] Humans are normally moral agents, with the exception of very small children and those with very severe mental disabilities.

Notice the asymmetry of what counts as a moral agent and what counts as a moral patient. My dog is a moral patient; he is not a moral agent. The same applies to very small children. It would be very wrong to ignore the welfare of a child. The (small) child cannot be blamed when they do something that would be wrong for an adult to do, like throw a rock at the dog since they do not yet have the cognitive abilities necessary for telling right from wrong.

6.3 Warm up

As warm up, Floridi notes that as humans have made moral progress, especially in the recent past, our understanding of who/what counts as a moral patient has enlarged. Many writers now include

- Future people
- Subjects of posthumous harms (harms to people who are dead)
- The environment / natural world
- Animals

in the class of moral patients.

Including a natural environment (e.g., a pristine forest) as a moral patient means that it merits moral consideration on its own right. This goes beyond saying that humans

---

17. If you believe that a volcano is a god which requires human sacrifice to keep it from erupting, then you probably believe the volcano is a moral agent.
18. We're probably using moral concepts as analogies. We are saying something like 'If a human behaved like that, they would be cruel'.

enjoy hiking in pristine forests and therefore we have reason to maintain them. It goes beyond saying that pollution of waterways has adverse effects on ecosystems which cause harms to humans. It means that we are to evaluate effects on the forest's interests alongside human interests.[19]

Similarly, saying animals are moral patients implies that we have reason to not be cruel to animals because it is wrong to do so. This contrasts with writers like Kant who thought that it is wrong to be cruel to animals because cruelty to animals hardens a person and makes her more likely to be cruel to people.

With that in mind, Floridi asks us to consider that maybe it is now time to enlarge the scope of what/who counts as a moral agent. In particular, maybe it's time to to treat artificial agents like computer programs as moral agents too.

Time to get crazy.

6.4 Starting points

Floridi wants us to approach this question with an adequately open mind. He thus proposes some ground-rules. He builds a framework of levels of analysis to situate and formalize those rules.[20] We'll ignore that framework and just focus on the principles he wants us to start with.

6.4.1 No anthropocentrism

According to Floridi, if we start from the assumption that everything within the moral realm is based on human beings, we've already closed the door to potentially important moral considerations coming from non-humans.

Thus if we want to be fully open-minded about algorithms, we must start by purifying ourselves and setting aside any anthropocentric (human-centered) biases.[21] He writes

---

19. If you're confused by how we would actually do this in practice, I'm right there with you. Most suggestions involve things like appointing a human advocate for the forest or just being absolutist and saying the interests of natural environments can't be traded off against human interests.

20. When you see 'LoA' in his paper, that's what he's talking about.

21. If you're seeing huge 'DANGER' signs and bright red blinking lights in your mind when you read this, I'm right there with you. To my mind, this move is the original sin of the argument and the place I would/will direct my attacks. But we'll get to that…

> Limiting the ethical discourse to individual agents hinders the development of a satisfactory investigation of distributed morality, a macroscopic and growing phenomenon of global moral actions and collective responsibilities resulting from the 'invisible hand' of systemic interactions among several agents at a local level. Insisting on the necessarily human-based nature of the agent means undermining the possibility of understanding another major transformation in the ethical field, the appearance of artificial agents (AAs) sufficiently informed, 'smart', autonomous and able to perform morally relevant actions independently of the human engineers who created them, causing 'artificial good' and 'artificial evil'. Both constraints can be eliminated by fully revising the concept of 'moral agent'. {Floridi:uy} p.3

If it helps you get in the swing of things, it may help to keep in mind that the human body, brain included, are (at least largely) machines. Doctors are glorified mechanics.

6.4.2 Observability

We're also going to conduct this discussion by defining everything in terms of things which can be observed. Since we can (in principle) see everything that's going on in a machine, it wouldn't be fair to demand that it have something non-observable. For example, if you thought that humans are moral agents because they have an immaterial spook that resides in them but no one can see (because it's immaterial), then you'd be biasing the inquiry against other possible entirely material agents such as machines.

In other words, if we are committed to not being anthropocentric, we will want to require observability.

Notice that this may be a bit important since most human psychology is unobservable (behavior is; mental states aren't). Thus everything we said above about mens rea looks like its going to be right out the window. That means Floridi is asking us for a heavy revision of our ordinary notion of responsibility. To be clear, that's not on its own an argument against him. But will be helpful to keep track of the costs.

6.4.3 Terminology

In order to avoid anthropocentrism in discussing agency, Floridi makes use of a picture loosely drawn from computer science. That entails some fancy terminology.

The systems we're interested in have an internal state which can be changed through various transition rules. They store information internally and have specific procedures for changing that information.

Let me illustrate this by a different version of the example he gives.[22] Consider a simple finite-state autonoma such as a vending machine or the gate at a (old school) parking garage which goes up when the appropriate amount of change has been put in.[23]

I'll write this out in (rough) Python. (Lines that begin with a '#' are comments for humans to explain what's going on and not part of the program; same for lines that are between two sets of 3 quotation marks: """This is a comment""")

First we define some variables that hold the information we need. This is the internal state of our machine.

```
# The amount of coins needed to raise the gate
required = 3

# The amount of coins that have been put in
paid = 0
```

Then we define some functions which actually do the work. These are the transition rules

```
def raise_gate():
    """Actually raises the gate"""
    # Turn on motor, etc, goes here

def coin_inserted(amount):
    """Function which gets called whenever someone
    puts a coin into the machine"""

    # Increase the stored value of what's been paid
```

22. His example is MENACE the tic-tac-toe learning matchbox machine. {Floridi:uy} pp. 8-10

23. Here's a relatively accessible video explanation of finite-state autonoma: https://www.youtube.com/watch?v=vhiiia1_hC4

```
# by the amount that just got put in.
paid += amount

# Now we check whether enough coins have been inserted
# by comparing the paid variable to the required variable
if paid >= required:
   # The customer has paid enough so
   # we call the function which raises the gate
   raise_gate()
         # Reset the paid variable for the next car
         paid = 0


# If not enough coins have been entered,
# nothing else happens. We wait for the function to be run
# again when another coin is inserted.
```

What happens is that every time a coin is inserted into the machine, we run the coin_inserted function. It adds the number of coins to the stored value paid (it changes its internal state). Then it checks whether enough coins have been inserted and if that's true, it opens the gate.[24]

This doesn't have to be an electronic process. The old version of these machines operated mechanically.

6.5 Agents

Let's start with what it is to be an agent in general, we can then understand moral agents as a proper subset of agents.[25] If we stick to criteria which must be observable and formulate them in ways which are not anthropocentric, Floridi claims we will find 3 characteristics of agents.[26]

---

24. Note that if you put more than the required number of coins in, you are screwed. There's no give_money back function. (For those of you learning python: you would want to write this using exceptions, that way you can also handle things like non-coins being inserted)
25. 'proper subset' means that all moral agents are agents but not all agents are moral agents.
26. [p.7]

6.5.1 Interactivity

First, an agent is interactive in that it can be affected by changes in its environment and when it can affect its environment.

> Interactivity means that the agent and its environment (can) act upon each other. Typical examples include input or output of a value, or simultaneous engagement of an action by both agent and patient —for example gravitational force between bodies {Floridi:uy} p.7

A rock can be affected by its environment. Hit it with a hammer and it changes. However, without some outside force acting upon it, the rock does not affect its environment.

Your computer is interactive. You type stuff in, it does stuff. Similarly, you and the Earth are interacting simultaneously through gravity. You are pulling up on the Earth just as hard as it is pulling down on you.[27]

A volcano doesn't really interact with its environment.The movement of magma which results in its eruption just is the thing that leads to the eruption. There's no intermediate step. If vulcanism was way different and volcanos erupted because their thirst for human sacrifice was unquenched this would be different, assuming that the volcano thinks "You know, I haven't had a human snack in a long time. I should get those villagers' attention…"

6.5.2 Independence

Second, an agent can change its internal state without direct response to interaction; it can perform internal transitions to change its state. Floridi calls this 'autonomy', but since we've been using that term in a very specific way, I'm going to call it independence.

> Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So an agent must have at least two states. This property imbues an agent with a certain degree of complexity and decoupled-ness from its environment. {Floridi:uy} p.7

27. Yep. Thanks general relativity.

Consider our friend the parking garage gate which goes up when the appropriate amount of change has been put in. It clearly isn't just responding to the input of a coin since it responds differently after the appropriate number of coins have been inserted. Thus it is independent in Floridi's sense.

6.5.3 Adaptability

Third, an agent is adaptable when it can modify the transition rules by which it changes its internal state.

> Adaptability means that the agent's interactions (can) change the transition rules by which it changes state. This property ensures that an agent might be viewed… as learning its own mode of operation in a way which depends critically on its experience. Note that if an agent's transition rules are stored as part of its internal state then adaptability follows from the other two conditions. {Floridi:uy} p.7

You demonstrate adaptability when you learn new things or change your mind. Suppose you didn't know that affirming the consequent is a logical fallacy.[28] Like all of us, you think the sentence 'If it has rained recently, the street is wet' is true. But when you look out the window upon a recently hosed sidewalk, you incorrectly conclude that it must have rained recently, and decide not to come to class. That's too bad. If you had come, you would've learned some logic and no longer been tempted to make bad inferences like that. You would see wet streets and look for other evidence that it has rained.

We leave our friend the parking garage gate behind with this criterion. Any time you've put in less than the required amount, it doesn't open; when you've put in the required amount, it opens. That never changes. On its own, it can never decide to hold your car hostage for 4 coins rather than 3. It can never decide to allow free exits to those who sympathetically portray garage gates in their teaching.[29]

Adaptability is one of the hallmarks of some artificial intelligence systems involving machine learning. Think of a convolutional neural network that detects cats in pictures. You start off with a bunch of pictures of cats labeled (true) and tons of other stuff labeled (false). The system starts with transition rules that are random —it guesses

---

28. In case it's been awhile: If you have a statement with the form `If p, then q` and learn that q is false, you can infer that p is false.
However, if you learn that q is true, you cannot infer that p is true. That's the fallacy of affirming the consequent. The example of the wet sidewalk illustrates why.
29. Obviously, I'm hedging my bets.

whether a picture contains a cat. On the basis of whether it was correct, it updates the rules by which it determines if something is a cat. Eventually, it gets really good at cat recognition through updating its own transition rules.

6.6 Moral agents

Let's suppose that we've now captured what it is to be an agent. Presumably, moral agents are a proper subset[30] of agents. Which ones are they? The ones that can do moral actions, duh. Or, as Floridi puts it, some actions are morally qualifiable. Moral agents are agents which can do morally qualifiable actions.

When is an action morally qualifiable? Floridi claims that

> An action is morally qualifiable if and only if it can cause moral good or evil.

This definition is allegedly neutral between consequentialist and 'intentionalist' theories. The moral goods and evils could be found, as consequentialists think, in states of the world. A world which contains 5 people suffering agony is worse than a world in which only 1 person suffers. Alternatively, they could be found, as non-consequentialists think, in the motives / character of persons. Blue intending to torture 5 people is worse than intending to give 5 people candy.

From there we can define a moral agent

> Agent S is a moral agent if and only if S is capable of morally qualifiable action.

That's all there is to it. Anything which meets the criteria for being an agent and which can perform morally qualifiable actions, will be a moral agent.

Any resistance to this can only be baseless anthropocentrism. To help us see this, Floridi gives an example of two entities H and W which are able:

---

30. That is, all moral agents are agents but not all agents are moral agents. Compare: diet coke is a proper subset of sodas.

> i) to respond to environmental stimuli — e.g. the presence of a patient in a hospital bed — by updating their states (interactivity), e.g. by recording some chosen variables concerning the patient's health….
> ii) to change their states according to their own transition rules and in a self governed way, independently of environmental stimuli (autonomy), e.g. by taking flexible decisions based on past and new information….
> iii) to change according to the environment the transition rules by which their states are changed (adaptability), e.g. by modifying past procedures to take into account successful and unsuccessful treatments of patients.

It seems H and W qualify as agents on our definition.

> Suppose that H kills the patient and W cures her. Their actions are moral actions. They both acted interactively, responding to the new situation they were dealing with, on the basis of the information at their disposal. They both acted [independently]: they could have taken different courses of actions, and in fact we may assume that they changed their behaviour several times in the course of the action, on the basis of new available information. They both acted adaptably: they were not simply following orders or predetermined instructions; on the contrary, they both had the possibility of changing the general heuristics that led them to take the decisions they took, and we may assume that they did take advantage of the available opportunities to improve their general behaviour. The answer seems rather straightforward: yes, they are both moral agents. There is only one problem: one is a human being, the other is an AA [Artificial Agent] ….So can you tell the difference? If you cannot, you will agree with us that the class of moral agents must include AAs like webbots." [12-13]

Remember, it's no good protesting that one is a human and the other a robot. Your resistance to this conclusion is just baseless pro-human prejudice. After all, you agreed to Floridi's setup where all the relevant concepts 'agent', 'moral action', et cetera are defined to be neutral between humans and non-humans.

It should be no surprise that if you drain our human moral concepts of anything specific to humans, you'll get a result that applies to non-human things. But we're getting ahead of ourselves. Before we start jumping up and down on Floridi's view, let's tease out its implications a bit more.

6.7 Are gatekeeper algorithms moral agents?

Let's see how Floridi's picture works with one place that artificial intelligence touches most of our lives by discussing what Zeynep Tufekci calls Gatekeeper algorithms. In

particular, let's focus on the algorithms which decide what posts you see on social media.[31]

Are gatekeeper algorithms moral agents? Let's run down the list.

Are they interactive? That is, does it do things in response to its environment? Yes. People do things on Instagram. The algorithm makes decisions about whom to show those things to. Then it shows them those things.

Are they independent? That is, do they change their internal state? Yep. The algorithm adds a new post to its queue of posts to make decisions about. It makes a decision about whom to show the posts too. Then it shows them the posts.

Is it adaptable? That is, can it change its own internal rules by which it changes internal state? Yes. This is the 'learning' part of 'machine learning'. The system has a complicated statistical model which predicts what posts will maximize engagement. It updates this model based on whether those predictions were correct.

Therefore, the gatekeeper algorithm is an agent.

Is it a moral agent? That is, can it cause moral goods and evils?

Well, there is a reasonable amount of research which suggests that people who heavily use social media suffer in psychological ways. They may have more depression and have lower self-esteem. Though the extent of the effect is pretty questionable.

There's no question that social media has contributed to violence and other moral evils.[32] Ideological violence seems to be socially contagious; seeing others of your ideological affinity commit violence increases the likelihood that you will do so. This was an integral part of the Islamic State's social media strategy.[33] Increasing the probability of innocent people being murdered seems pretty clearly a moral evil.

---

31. {Anonymous:SbzjS1LM}
32. [ToDo Add ref to mob violence case in India]
33. [ToDo ref]

But is the algorithm the relevant agent? Well, at least at Facebook, that's literally the official story.[34] For example, according to the company itself, Facebook's algorithm neglected to prevent a terrorist from live-streaming the murder of innocent people at a New Zealand mosque

> The members of Congress who gathered for a closed-door briefing had lots of questions for Brian Fishman, Facebook's policy director for counterterrorism. One of the biggest: Why didn't Facebook's counter-terror algorithms—which it rolled out nearly two years ago—take down the video as soon as it was up? Fishman's answer, according to a committee staffer in the room: The video was not "particularly gruesome." A second source briefed on the meeting added that Fishman said there was "not enough gore" in the video for the algorithm to catch it. Members pushed back against Fishman's defense. One member of Congress said the video was so violent, it looked like footage from Call of Duty. Another, Missouri Democrat Rep. Emanuel Cleaver, told The Daily Beast that Fishman's answer "triggered something inside me." "'You mean we have all this technology and we can't pick up gore?'" Cleaver said he told Fishman. "'How many heads must explode before they pick it up? Facebook didn't create darkness, but darkness does live in Facebook.'"[35]

Therefore, the Facebook gatekeeper algorithm is a moral agent.

Let's try to figure out what that means.

6.8 Moral agency without responsibility

Let's consider the implications of what Floridi has (allegedly) shown. He writes

---

34. https://www.fastcompany.com/40475913/facebook-and-google-apologies-for-fake-news-ignore-the-system-itself
https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters
https://www.wired.com/story/facebook-can-absolutely-control-its-algorithm/
https://www.ibtimes.co.uk/facebook-blames-spam-algorithm-blocking-links-wikileaks-dnc-email-leaks-1572488

Refs from https://boingboing.net/2019/04/10/once-again-facebook-blames-an.html
35. https://www.thedailybeast.com/facebook-tells-congress-new-zealand-shooting-video-wasnt-gruesome-enough-to-flag

> agents (including human agents) should be evaluated as moral if they do play the 'moral game'. Whether they mean to play it, or they know that they are playing it, is relevant only at a second stage, when what we want to know is whether they are morally responsible for their moral actions. [14]

Thus we should separate the question

> Is x a moral agent?

from

> Is x morally responsible for its actions?

The class of things which are moral agents is larger than the class of things which can be morally responsible for their actions.

He claims that prescriptive (normative?) discourse is larger than responsibility attribution. Indeed, we can morally evaluate actions in beings incapable of responsibility

> Good parents…engage in moral-evaluation practices when interacting with their children even at an age when the latter are not responsible agents, and this is not only perfectly acceptable but something to be expected. This means that they identify them as moral sources of moral action, although as moral agents they are not yet subject to the process of moral evaluation. [15]

Similarly, a search-and-rescue dog can save a person's life; my dog may decide to give me the snap and injure me. Those are moral goods and bads. So we can call them moral agents, without saying that they are morally responsible for their acts. The author of a SEP article elaborates on something Floridi mentions in passing about dogs

> Dogs can be the cause of a morally charged action, like damaging property or helping to save a person's life, as in the case of search-and-rescue dogs. We can identify them as moral agents even though we generally do not hold them morally responsible, according to Floridi and Sanders: they are the source of a moral action and can be held morally accountable by correcting or punishing them. [36]

That last part is key. A moral agent which is incapable of moral responsibility is still subject to moral accountability. What is that?

> The whole conceptual vocabulary of 'responsibility' and its cognate terms is completely soaked with anthropocentrism. This is quite natural and understandable, but the fact can provide at most a heuristic hint, certainly not an argument. The anthropocentrism is justified by the fact that the vocabulary is

36. https://plato.stanford.edu/entries/computing-responsibility

> geared to psychological and educational needs, when not to religious purposes. We praise and blame in view of behavioural purposes and perhaps a better life and afterlife. Yet this says nothing about whether or not an agent is the source of morally charged action. Consider the opposite case. Since AA [Artificial Agents] lack a psychological component, we do not blame AAs, for example, but, given the appropriate circumstances, we can rightly consider them sources of evils, and legitimately re-engineer them to make sure they no longer cause evil. We are not punishing them, anymore than one punishes a river when building higher banks to avoid a flood. But the fact that we do not 're-engineer' people does not say anything about the possibility of people acting in the same way as AAs [14]
>
> The point raised by the objection is that agents are moral agents only if they are responsible in the sense of being prescriptively assessable in principle. An agent x is a moral agent only if x can in principle be put on trial. Now that this much has been clarified, the immediate impression is that the objection is merely confusing the identification of x as a moral agent with the evaluation of x as a morally responsible agent. Surely there is a difference between being able to say who or what is the moral source of the moral action in question and being able to evaluate prescriptively whether and how far the moral source so identified is also morally responsible for that action. Well, that immediate impression is indeed wrong. There is no confusion. Equating identification and evaluation is actually a shortcut. The objection is saying that identity (as a moral agent) without responsibility (as a moral agent) is empty, so we may as well save ourselves the bother of all these distinctions and speak only of morally responsible agents and moral agents as synonymous. And here is the real mistake, because now the objection has finally shown its fundamental presupposition: that we should reduce all prescriptive discourse to responsibility analysis. But this is an unacceptable assumption, a juridical fallacy. There is plenty of room for prescriptive discourse that is independent of responsibility-assignment and hence requires a clear identification of moral agents. [15]

What problem does this solve?

> Our more radical and extensive view is supported by the range of difficulties which in practice confronts the traditional view: software is largely constructed by teams; management decisions may be at least as important as programming decisions; requirements and specification documents play a large part in the resulting code; although the accuracy of code is dependent on those responsible for testing it, much software relies on 'off the shelf' components whose

> provenance and validity may be uncertain; moreover, working software is the result of maintenance over its lifetime and so not just of its originators…. Such complications may point to an organisation (perhaps itself an agent) being held accountable. But sometimes: automated tools are employed in construction of much software; the efficacy of software may depend on extra-functional features like its interface and even on system traffic; software running on a system can interact in unforeseeable ways; software may now be downloaded at the click of an icon in such a way that the user has no access to the code and its provenance with the resulting execution of anonymous software; software may be probabilistic; adaptive; or may be itself the result of a program (in the simplest case a compiler, but also genetic code). All these matters pose insurmountable difficulties for the traditional and now rather outdated view that a human can be found responsible for certain kinds of software and even hardware. Fortunately, the view of this paper offers a solution at the 'cost' of expanding the definition of morally-charged agent.

6.9 Objections

Now that we've got Floridi's picture on the table and seen what it's supposed to help with, we can turn to assessing it. Should we buy what Floridi is selling? I think not.

To begin, we really need to go back and think through his starting point. Remember, we were just supposed to assume a non-anthropocentric and entirely observable notion of moral agency. But what if there are crucial components to our ordinary conception of agency which this approach has just waived away? If that's the case, the foundations of his account are built on sand.

So what's missing? Think of an accountant who decides to go along with the CEO's demands and report false numbers to the Board. She gets caught and is punished for the fraud. What would we normally say about her? Probably that she made the wrong choice. That she paid too much attention to the CEO and not enough attention to her ethical obligations.

Now think way back to when we first talked about autonomy. Remember the example of the person who has a spasm and smacks the person sitting next to them in the face? Because the action was beyond her control and she didn't intend to hit the person, we said she wasn't responsible.

The point is that any thinking about human agency involves decision-making, i.e., the use of some form of reason. The unethical accountant is blameworthy because she made the wrong decision. The spasmodic smacker is blameless because she made no decision.

With that in mind, let's train our sights on Floridi's account of moral agency. I'm not going to go after the specifics of the view. I'm going to target the approach which supports it.

6.9.1 Contra observables

Already we have some tension with Floridi's requirement that agency be based on observables. Consider our corrupt accountant. What would we see as she commits her misdeeds? Well, she takes a call, hangs up the phone and sits at her desk for awhile staring at the wall. Then she types some stuff in Excel, makes a fantastic PowerPoint, and heads upstairs to the boardroom. Our spasmodic sits quietly on the train and then suddenly smacks the person next to her. The unobservable things going on inside their heads are the crucial factor in our judgments of their moral responsibility.

Floridi does consider an objection that runs along similar lines. He imagines his opponent claiming that

> To be a moral agent, the AA must relate itself to its actions in some more profound way, involving meaning, wishing or wanting to act in a certain way, and being epistemically aware of its behavior. [13]

His response, having set out an account of agency is straightforward

> agents (including human agents) should be evaluated as moral if they do play the 'moral game'. [14]

That is, we've agreed that observability is a criterion of any account of moral responsibility, so the opponent loses.

But wait. Moral responsibility is a concept humans have had for a very long time. Every culture has some version of it. We use it unthinkingly in evaluating whether the accountant made a bad decision. If it has unobservable conditions, so what? I don't know your motives. They are hidden in your head. They still matter for determining what you are responsible for. Indeed, this was the whole point of discussing mens rea above. Morally defective mental states are a requirement of responsibility.

Thus let me bring out a few things which are very standardly pre-conditions of moral responsibility. For one, you need to be able to use reason. Reason has standards, namely

logic.[37] To use logic, you need to have a concept of beliefs/claims being true or false. Apparently, children younger than 2[?] lack the concept of truth and falsity. They cannot apply the concept to their situation. If that's true, a very small child cannot be morally responsible since she does not have the relevant capacity for reason.

Does an algorithm have the capacity for reason? On the one hand, it is certainly built out of logic. All non-quantum computers are at the end of the day just a lot of logic circuits. But being able to make logical decisions does not entail having the concept of logic or the concept of decisions. Arguably a computer does not have the concept of true or false;

In this vein, recall what Floridi's says about small children

> Good parents…engage in moral-evaluation practices when interacting with their children even at an age when the latter are not responsible agents, and this is not only perfectly acceptable but something to be expected. This means that they identify them as moral sources of moral action, although as moral agents they are not yet subject to the process of moral evaluation. [15]

Does it? I think not. Or, at least, I can think of two possible alternative explanations of what's going on that don't require the child to be a moral agent.

First, because they are human children they will eventually become sources of moral action. Thus we treat them *as though* they are moral agents to help them learn to become moral agents. Indeed, since the relevant psychological capacities likely come in gradually, treating them as though they are agents helps them 'grow into' actual agency.

Second, human brains are lazy. We are used to relating to other adult humans as moral agents. Thus our default way of relating to other creatures is as though they are moral agents too. We treat the child as an agent because we are used to treating people that way; if we stop and think, we catch ourselves making unreasonable judgments about infants.[38] Think also of how we find ourselves treating pets as agents. You might think

---

37. That's not to say people need to be able to formulate these standards or explicitly use the rules you learn in logic class.

38. As one of my friends once told me about her infant "He keeps dropping his pacifier and then gets upset. I find myself thinking things like 'if you want it, why'd you drop it?' and then I think I'm crazy. He's a baby, what am I talking about."

this is what's going on if you've ever found yourself telling the cat, who just came in 30 seconds ago and is now demanding to go out, to make up its damn mind.

6.9.2 Contra anthropocentrism

The other ground rule which Floridi raises is the specter of anthropocentrism. He claims that if we set up the game to exclude non-human agents, we won't be able to take seriously the possibility of artificial agents. That's fine. But it is quite another thing to drain moral responsibility of everything that makes it distinctive and a subject of concern.

Notice the way the paper ends with him trying to explain what it means for a machine to be a moral agent that creates a moral harm. We might, for example, do things to it. It may be reprogrammed or taken out of service morally responsible.

But there is nothing that the machine should do to make up for the moral harm it causes. The company which owns it may have obligations. The developers who work on it may have obligations. But it doesn't even make sense to say that the gatekeeper algorithm in Internet isolation should do something different. All we can say about it is that it does something bad by not showing our victim's posts to others. But we have to say that the programmers are the ones who should revise it so that it no longer creates the harm.

Any form of actual moral agency identifies the agent as the wrongdoer (or good doer), sure. But it also means that that agent ought to act differently. It means that agent ought to feel bad for what they've done. It means that agent should make amends. In other words, there is more than one kind of moral action that is tied to moral agency. If a non-human machine can only do one kind of action, it cannot be a full moral agent.

If this is anthropocentrism, then so what? Ethics and morality are features of human life. We had it first. If the robots want it too, they need to be more like us; we needn't conform our concepts to them.[39]

6.9.3 The asymmetry

Finally, a point about this whole approach. Recall that to warm us up, Floridi pointed out that we've already expanded the scope of moral patients in order to make us more

---

39. Note that what I'm saying here does not affect attempts to extend moral patient status to, say, animals. Those arguments are based on animals having interests.

receptive to expanding the scope of moral agents. This claim isn't crucial to his paper. He could give it up and still do everything he wants. But it's worth thinking about since approaching the history of moral progress in the way he suggests misses a lot.

Maybe there's a good reason we've been slower to expand the scope of moral agents than moral patients. We've expanded moral patient-hood to things whose interests we recognize need to be weighed up alongside human interests. Future generations and animals have (or will have) interests like ours and we've recognized that they ought to be weighed alongside ours. The environment is a bit different; it is not widely accepted that the environment could have interests independent of human and animal interests (in part because it is still not entirely clear how to understand this claim).

Indeed, including other beings as moral patients does not entail granting them equal status as humans. It is wrong to torture a mouse for your amusement. But if you could only save a human child or a mouse from a fire, it would be wrong to choose the mouse.