Free Will Debate Survey

Background. We want to know how to make sure that our algorithm does not threaten the users' autonomy. To that end, our algorithm needs to be able to do two things: (1) recognizes the users' (potential) autonomous choices; and (2) knows what the algorithm must do with the info obtained in (1) so what it does doesn't count as threatening the users' autonomy. To make sure that our algorithm can do (1), we need to understand what it is that we want our algorithm to recognize. That is, we need to understand what a free choice even is. The following is an overview of the philosophical literature on the nature of free will. I divide the literature into three periods: the Classical Period, 1960s-70s, and the Contemporary Period.

1. Classical Period

The philosophical community before 1960s is rather dominated by **classical compatibilism** (defended in Hobbes (1651), Hume (1748), Schlick (1939), Ayer (1954)). Compatibilism is the view that it is not incompatible to say that (i) our actions are causally determined *and* (ii) some of our actions are results of free choices. According to the *classical* compatibilists, compatibilism is true for two major reasons:

[a] An action X is autonomous if and only if the following two conditions are met:

(i) X is not under coercion or duress.
(ii) If the agent had chosen otherwise, he could have done otherwise (than X).

(i) + (ii) basically says an action is free if and only if it is produced by what the agent chooses. Even if causal determinism were true (i.e., a complete description of a moment in the past plus the laws of nature jointly determines what happens in the future, including what we do), both (i) and (ii) can still be true. (Notice that (ii) is a subjunctive conditional.)

[b] Events that are not determined are incompatible with free choice. According to classical compatibilists, if an event is not determined, it is randomly so. A random event is not a choice at all. So, autonomy has to be compatible with determinism.

Comment. *As a theory of autonomy, classical compatibilism has very few defenders nowadays (e.g., Hyman 2014). That is because there are many counterexamples that are hard to explain away. E.g., a brainwashed person can perform action that is the product of his choices but still that shouldn't be considered as free actions. The same goes for people with phobias. Doing what one chooses is not sufficient for autonomy. (See Chisholm (1967), Lehrer (1968))*

2. 1960s-70s

The publication of three arguments in the 60s-70s generated a big literature. Two of the arguments are against compatibilism, one for.

[a] The Consequence Argument.

This is the first rigorously articulated argument for **incompatibilism**. Incompatibilism says that causal determinism is incompatible with autonomy. The Consequence Argument was first developed by Ginet (1966; 1990), then by Wiggins (1973)… But it is the version developed in Peter van Inwagen (1975; 1983) that has drawn the most attention. Stripped to the core, the argument can be presented as follows:[1]

Let P be a statement that completely describe a random moment in the past, L be a complete description of the laws of nature, and X be a statement about the present.

> Premise 1: It is not a choice of anyone that P & L
> Premise 2: It is not a choice of anyone that P & L entails X
> Conclusion: It is not a choice of anyone that X

Note that (i) we can substitute X with any statement that describes the actions we perform, and (ii) determinism entails premise 2. (Premise 2's plausibility is ultimately based on the thought that it is not anyone's choice that determinism is the case.)[2] Hence, the Consequence Argument says: *if* we accept determinism and premise 1, we need to accept that there are no free choice. This shows that determinism and free choices are *incompatible*.

Objections come from two directions: (1) challenging premise 1 (Lewis); (2) denying that assuming premise 1 and premise 2 is enough to demand the conclusion, i.e., denying that if Q logically follows from P, 'it is not a choice of anyone that Q' follows from 'it is not a choice of anyone that P' (Slote (1982), McKenna (2008)).

---

[1] This might be unnecessary, but just in case this is something that is not common outside my discipline. In philosophy, we always try to decompose a piece of reasoning into premises and conclusion to make it easier to scrutinize whether the piece of reasoning is really airtight, for we are too easily misled by our ideologies, biases, and rhetorics. The idea is simply that, all reasoning must start from a set of assumptions, i.e., premises, and end with one conclusion. A good reasoning is one that (1) the premises are all true and (2) accepting the premises and denying the conclusion results in inconsistency. In this case, the incompatibilists do not mean to use the Consequence Argument to show that we do not have free choices. The point is just to show that, ***if*** we assume determinism (premise 2), we need to conclude that way. It seems that most incompatibilists in fact accept that free choice is a real phenomenon (and consider that evidence for rejecting determinism instead). It's perhaps also worth mentioning that incompatibilism is compatible with the idea that there are deterministic systems in the world.

[2] This is a simplified version. It is in fact a bit more complicated. For determinism to plausibly entail premise 2, according to Van Inwagen, we also need two extra assumptions: (1) if determinism is true, it is *necessarily* true; and (2) if it is necessary that p, it is not anyone's choice that p. So, the most prominent version the Consequence Argument leaves open the possibility that a *contingently* deterministic world contains free choices.

Comment. *It is worth noting that, if we understood the phrase 'it is (not) a choice of…' according to the classical compatibilism's subjunctive analysis of autonomy, the premises of the Consequence Argument fail to force the conclusion upon us. But as I have mentioned, classical compatibilism's prospect is plagued by difficult counterexamples. It is also worth noting that the Consequence Argument does not offer us a positive proposal for analyzing autonomy. All it does is tell us what a proper analysis should end up look like: it should be incompatible with determinism.*

[b] The Manipulation Argument

The argument in its most recognizable form is proposed by Taylor (1974) and in a more sophisticated form recently by Pereboom (2004). Again, stripped to the core, the idea is simple.

A compatibilist must offer an analysis of what free will consists in. Suppose, according to the compatibilists, T is the feature an action must instantiate that makes it autonomous. Since this is a compatibilist account, an action must be able to instantiate T and be autonomous even if it is causally determined that the action has T. So, it is possible that a manipulator can have it arranged and determined that I perform an action with the property T. If my actions are so manipulated, the are not free.

Now that in itself does not yet pose any problem for compatibilism because the view only says that an action can be determined and be autonomous. It does not say that whenever an action is determined, it is autonomous. So there is nothing inconsistent for a compatibilist to say that my action is not free when I am determined by a manipulator to perform an action with the feature T.

But according to the Manipulation Argument, the compatibilists need to give us an account as to why I am not free in the manipulation case but I am free when the 'manipulator' turns out to be my past plus laws of nature. If the compatibilists cannot highlight a relevant and illuminating difference between the two cases to help make sense of why my action is free in one case but not the other case, the view just becomes problematically arbitrary. That is the Manipulation Argument.

Comment. *This argument leads to a division in the contemporary literature between two camps: those who take historical factors to be crucial in the proper analysis of free will (e.g., Pereboom, Bratman, etc.) and those who don't (e.g., Frankfurt, Watson, Wolf, Nelkin, etc.). The former camp believes that a proper analysis of autonomy must include certain information about the causal history of one's action and the mechanism that brings about the action. This info will screen out the manipulated actions as not free. The other camp thinks that whether a person's action is free has nothing to do with the person's history. The Manipulation Argument has to be dealt with by something else. This division between historical vs. a-historical accounts of autonomy cuts across the compatibilism vs. incompatibilism divide. Curiously, as far as I can tell, there aren't much direct engagement between the two camps. People in the historical camp just go on to*

develop historical accounts of free will and the people in the a-historical camp just mind their own business and develop a-historical theories of free will.

[c] Neo-Compatibilism

Frankfurt (1969) offers an argument to show that (i) autonomy should not be analyzed in terms of *actually* being able to do otherwise;[3] and (ii) determinism and autonomy must be compatible. The argument for (i) is based on a thought experiment. Here is a rough version of it:

> Suppose there are two persons *Big Brother* and *Derek*. Derek is about to make a choice whether to do X or not. What Derek does not know is that Big Brother installed a chip in his brain. The chip does two things. First, it can detect prior signs that indicate that Derek is *about to* choose not to do X. Second, it is connected to a remote control that allows Big Brother to press a button to cause Derek to choose to do X. Here is Big Brother's plan. If Derek chooses to do X, he won't even touch the remote control. If the chip picks up signals that Derek is about to not do X, Big Brother will step in and use to remote to cause Derek to do X. All this would happen without Derek's knowledge. As the situation unfolds, Derek chooses to do X without Big Brother's interference.

Apparently, Derek's action is autonomous. That is because Big Brother didn't do anything and the chip does not participate in the performance of the action. Derek does X as if Big Brother and chip don't exist. If there are autonomous actions at all, that seems to be a paradigm case of an autonomous action. A good analysis of autonomy should be able to classify this case properly as autonomous.

Any account of autonomy that requires the agent to be able to actually do otherwise would misclassify Derek's action as not autonomous. That is because, in the situation, Derek is freely doing X but he is not capable of doing otherwise. Conclusion: a good analysis of autonomy should not require the agent to be able to do otherwise for an action to be free.

Based on (i), Frankfurt offers the following reason for (ii):

> Premise 1. Free choice and determinism are incompatible *only if* free choice consists of the ability to do otherwise.
> Premise 2. Free choice does not consist of the ability to do otherwise.
> Conclusion. Free choice and determinism are compatible.

Compatibilism that is based on the kind of thought experiment I described (called Frankfurt Cases in the literature) is called neo-compatibilism. Neo-compatibilism, unlike classical

---

[3] The qualification 'actually' is to keep a distance from the classical compatibilists' analysis of autonomy as *counterfactual* ability to do otherwise (i.e., if so and so *were* different, the agent *would have done* otherwise). What is being denied by Frankfurt is the analysis of autonomy as the ability to do otherwise in the *actual* situation.

compatibilism, does not rely on the subjunctive analysis of autonomy. Just like the Consequence Argument, the Frankfurt Cases do not give us a positive theory of free will. They only promise to show that whatever analysis of free choice we end up endorsing, it should not be incompatible with determinism.

Comment. *Based on my literature review, there are several objections against neo-compatibilism. They are all objections purporting to show, in one way or another, the Frankfurt Cases do not show what neo-compatibilists think they do. My estimation is, the most important objection is called the Dilemma Defense. Roughly put, the Dilemma Defense is an attempt to show that using the Frankfurt Cases to show incompatibilism is true tacitly requires the assumption of incompatibilism to begin with. (See Kane (1985; 1996); Widerker (1995))*

3. Contemporary

During 1960s-70s, arguments for or against compatibilism do not rely on any positive account of free will. (This is meant to be a virtue of the arguments. The argument is much stronger if it does not need to rely on assuming a very specific theory of free will.) Following the publication of these three arguments (and the various versions of them), a division of labor appears in the literature.

On the one hand, there are continuous heated debates to this day regarding the validity of the Consequence Argument, Neo-Compatibilism, and the Manipulation Argument. On the other hand, many philosophers try to move past the more abstract debates and go on to propose various concrete theories of autonomy. Some accounts are developed based on the assumption that neo-compatibilism is right (and the objections can eventually be overcome). Others assume that incompatibilism is right and develop accounts of free will that is incompatible with determinism. Some assume that the Manipulation Argument shows that certain historical elements must be in the correct analysis of free will and goes on to develop theories that have a historical dimension. Others maintain that a proper theory of autonomy has to be a-historical. Here are some of the most influential ones.

A. Compatibilist Accounts

(1) Mesh Theories. According to the mesh theories, an action is autonomous if it is based on a coherent integration of the agent's psychological states. Different versions of mesh theories propose different kind of mental integration as the mark of autonomy.

(i) *Frankfurt's (1971) Higher-Order Volition Theory.*

We need to distinguish between first order and second order desires. When I want a piece of pizza, that is a first order desire. When I want to act according to the first order desire, I have a second order desire on top of the first order desire. It is a desire about another desire.

Our first and second order desires can come apart. Frankfurt uses the example of an unwilling drug addict to illustrate the possibility. When a drug addict wants to do drugs but do not want to be the kind of person who wants to do drugs, the drug addict's first order desire (to do drugs) does not cohere with his second order desire (to not to desire drugs).

According to Frankfurt's account of autonomy, to act freely is to act according to the lower order desires that do not conflict with our higher order desires. In the drug addict example, he is not acting freely when he does drugs, even if he has the first order desire to do drugs. Instead, the addict is being 'enslaved' by his first order addictive desires. On the contrary, a drug addict who wants to do drugs and have no higher order desires against having the desire for drugs is free when he does drugs.

Frankfurt's theory of autonomy does not seem to be very widely accepted, but it is the first mesh theory every articulated. All other versions of mesh theories are developed by a contrast with Frankfurt's view. Objections to Frankfurt's view: Watson (1975), Haji (1998; 2002), Mele (1995), Pereboom (1995; 2001)

(ii) *Watson's Evaluative Judgment Theory.*

Watson argues that Frankfurt's account has trouble dealing with conflicting higher order desires. (If I have conflicting higher order desires, then whatever I do would fail to cohere with my higher order desires and nothing I do would be autonomous.) Instead of analyzing autonomous actions in terms of coherence with higher order desires, Watson (1975) proposes that we should understand autonomy in terms of our desires' coherence with our *evaluative judgments*, not higher order desires. Judgments and desires are two different kinds of mental states. Evaluative judgments are judgments about values. In the case of unwilling drug addict, his action is not autonomous, according to Watson's account, because acting on his desires to do drugs go against his own evaluative judgment against doing drugs. Alfred Mele (1995; 2007) further develops this approach to the mesh theory as a theory of self-control or self-governance.

(iii) *Bratman's Planning Theory.*

Both Frankfurt's and Watson's views are a-historical. Bratman (2007), however, develops a version of the mesh theory that gives a historical aspect to autonomy. Depending on ones view on the Manipulation Argument, this could be a reason to prefer Bratman's version of mesh theory. Very roughly put, an autonomous action is brought about but desires that cohere with the agent's *long term plan* (instead of higher order desires or evaluative judgments) for himself developed over time.

(2) Reason Responsive Theories. According to the Reason Responsive Theories, an action is autonomous if the agent's action is responsive to (i.e., receptive of and reactive to) reasons. Different versions of the view spell out the idea of reason responsiveness in slightly different ways.

(i) *Wolf's Disjunctive Theory.*

Wolf (1987) is moved by the neo-compatibilist argument. But she thinks that there is an important aspect of autonomy that all other theories fail to capture. She thinks that a person's action is not free if the person lack the capacity to act according to what is *good* (i.e., positive values) and *true* (i.e., facts). A delusional person is, for example, not capable of performing autonomous action, according to Wolf's view. This view calls for a disjunctive treatment of autonomy in the following sense: if I am acting for good reasons and based on accurate information about facts, my actions are autonomous; but if I act for bad reasons or based on false conception of the world, whether my actions are autonomous will depends on whether I could have acted according to good reasons and true beliefs — the ability to do otherwise is necessary in the latter but not in the former case, hence *disjunctive*. Recently, the view is developed and defended by Dana Nelkin (2011).

Comment. *One of the main task in defending this view is to handle the Frankfurt cases because the view requires the ability to do otherwise for free 'bad' actions. That is in conflict with Frankfurt's contention that ability to do otherwise is not needed in all cases. Nelkin (2011) borrows resources from Alternative Choice Theories (below). I find the response to Frankfurt cases quite ingenious, but I'm still not sure it works.*

(ii) *Strawsonian Normative Theories*.

This is a very influential view that began with Strawson (1969) and continued in one way or another by Wallace (1994), Russell (1995), and Scanlon (1998). This is a very distinct approach of compatibilism. And I'm not confident that I fully understand the view. The crux of the proposal, I take it, consists of two points:

> [a] Autonomy is about whether an agent is *responsible* for an action. According to defenders of this approach, 'being responsible' is essentially an ethical notion.
> [b] Strawson and his followers are anti-realists about ethics. They do not think that ethical goodness or badness are objective feature. Instead, they think that they are constituted by the typical moral sentiments of the members in a community.

Combining [a] and [b], saying that an action is autonomous and that the agent is responsible for it is not to attribute some objective feature to an action. Instead, whether an agent is responsible for an action depends on the reactive attitudes or moral sentiments of people in the agent's community (e.g., blame, resentment, praise, etc.) The Strawsonian view reverts the explanatory order of things: instead of first asking whether a person is *in fact* responsible for his actions and then use some ethical principle to decide what ethical responses are suitable, it argues that it is a mistake to think that there is a question about whether an agent is as a matter of fact responsible before the ethical questions about how we should react to the action. Whether a person's action is free, instead, depends on how the community members' ethical reactions to it. Ethical reactions

justify claims about autonomy, not the other way round. As a result, determinism becomes irrelevant in deciding whether people are autonomous.

This is a rough sketch. Different proponents of this approach further articulate autonomy's dependence on society in a different way.

Comment. *I'm trying to summarize the view as much as I can. It asks for a big revision in the way we think about the question of free will. And I admit that I have a hard time fully wrapping my head around.*

(3) Alternative Choice Theories. Most contemporary compatibilists are convinced that the Frankfurt Cases provide strong evidence for compatibilism. Hence, most compatibilists think that the proper theory of autonomy should not require the agent's ability to do otherwise. But recently, borrowing from philosophical discussions about the nature of *dispositions* (e.g., solubility of salt is a dispositional feature of salt), there is a revival of the idea that the ability to do otherwise is needed for autonomy and that ability is compatible with determinism.

(i) *Vihvelin's Masked Disposition Theory*.

This view is developed in Vhvelin (2013). It is, to a certain extent, an attempt to revive classical compatibilism. The focal point of developing this view is to handle the Frankfurt cases and find a way to show that the person under the watch of Big Brother (in the example I used to explain the Frankfurt cases) in fact has ability of do otherwise. Vihvelin relies on the idea of masked disposition to do so. Table salt is soluble. That is a dispositional property that salt has even when it is not actually put in water.

A dispositional property can be masked. For example, if I put some salt in a very well secured and sealed jar to prevent it from even dissolving, the dispositional feature of salt is thereby masked. But that does not mean I have taken solubility from salt. The salt remains soluble even though it is prevented from actually dissolving.

Here is how Vihvelin handles the ability to do otherwise. When a person do X and we say he could have done otherwise, all we mean is that the person has the dispositional property to do something else in the same sense that salt is soluble even if it is not dissolved actually. In the Frankfurt cases, the presence of Big Brother and the chip in Derek's brain only masks Derek's ability to do otherwise by preventing the disposition to do otherwise from manifesting. Just like putting salt in a jar does not rob salt its solubility, having the chip *standing by* does not rob Derek's ability to do otherwise. Hence, Frankfurt cases are not cases where a person acts autonomously without ability to do otherwise.

Comment. *This is, I think an insightful approach, but I am skeptical that it will be successful in handling all the potential problems of the view. The most important problem of all is this. It is hard to see how the view can avoid implying that, even when Big Brother activates the chip in*

Derek's brain, he is just preventing Derek's disposition to do otherwise from manifesting instead of taking away Derek's ability to do otherwise. If so, the theory has a hard time correctly classifying the difference between Derek doing X on his own and Derek doing X because Big Brother activates the chip.

B. Incompatibilist Accounts

Incompatibilists are people who believe that free choice and determinism are incompatible. The view that incompatibilism is true and free choice is a real phenomenon is called libertarianism (absolutely nothing to do with political libertarianism). An incompatibilist needs not be a libertarian because a person can think that free choice is incompatible with determinism and believe that free choice does not exist. But my survey focuses on people who believe that free choice is a real phenomenon. There are three major theories of libertarianism.

(1) Non-Causal Theory. The most prominent defender of this approach is Ginet (1990). According to the non-causal theory, an action is free when it is not caused. A person can decide to take his desires, beliefs, values, etc. into consideration. But when he freely takes any action, the action is not caused by those desires, beliefs, or values. The major issue with defending this view is to articulate in what sense an uncaused even can be *up to the agent* instead of being just a random event (Ginet 2007). This is easier said than done.

(2) Event-Causal Theory. Judging from what I have read, this is the most prominent view among libertarians. This view requires the idea of causation does not need to be deterministic; instead, there are probabilistic causations. Say I perform an action X. X is a freely chosen action if and only if there is a causal chain that traces from X back to the agent's action Y that is probabilistically caused. (X may be identical to Y.) The probabilistically caused actions like Y help constitute the personality/character of the agent that fixes the agent's later actions. As long as these later actions flow from the agent's own character, those actions are free. This view is developed and defended more prominently by Kane (2000). The most pressing problem in developing this view is to find a way to make sense of the difference between a random event that happens to a person that he has no control and the probabilistically caused actions the theory proposes.

(3) Agent-Causation Theory. This is perhaps the most exotic theory of autonomy in the literature. First of all, this view rejects the non-causal theory because, without the notion of causation, it is hopeless trying to make sense of the fact that we are responsible for the actions we perform. But at the same time, this view denies that probabilistic causation is relevant in shedding light on autonomy, which consists of control. Probabilistic events are like coin flipping. There is no relevant sense of control involved.

To have free actions caused by not determined, advocates of the agent causation theory suggests that we need a better understanding of the notion of causation. We speak of causal relation. Relation needs relata. Typically, causal relations are thought to link up *events* as relata. (Linguistically, an event is expressed by complete sentences.) Chisholm (1964) and O'Connor

(2009) argues that there is another kind of causal relation that does not connect event to event. Instead, this kind of causal relation connect an entity to an event, where the entity is an agent and the event is the action performed freely by the agent. This is called agent causation.

Whereas an event can cause another event and (if there is agent causation) an agent can cause an event, there is no event causing an agent. Hence, a free action is a case that is an agent causing an event without anything else causing the agent. That is why a free action is when an agent is the ultimate source of what happens. An alleged major advantage of introducing agent causation is that this is the only view of autonomy that has the theoretical resources to distinguish an action which is an event brought about by the agent from an event that just happen to a person. (E.g., me having a desire is something that happens to me, and the fact that this desire goes on to cause something else is also something that happens to me. Without agent causation, it seems that the way we think about actions will leave no room for the agent to play any active role. That seems to be a problem.)