Phil 134/ 234 G
Moral Psych: S 21

Prof. Aaron Z. Zimmerman
April 26, 2021

Handout #3: Sober and Wilson on the Nature of Altruism

I. PSYCHOLOGICAL EGOISM DEFINED

Psychological Egoism (S&W def.): All of our *ultimate* desires are self-directed. Whenever we want some other person to do well (or badly) we have that desire *only* because: (a) we have a self-directed desire for some personal benefit x, and (b) we believe that if that other person does well (or badly) we will get x (or are more likely to do so).

Question: Is the truth of psychological egoism compatible with the existence of altruism?

> **S&W's Claim**: For every apparently non-egoistically motivated action yet imagined, an egoist can come up with an egoistic set of motives that would equally well explain the agent's actions.

A deeper question: How do we evaluate competing claims about what an agent wants and believes? How do we evaluate competing claims about which beliefs and desires led an agent to act in the way that she did?

> Underdetermination of theory by evidence: No theoretical claim C, is logically entailed by a set of observations O. For any set of observations O, there will be mutually incompatible theories each of which is "commensurate" (or compatible) with the data provided by the theory.

A Central Issue in the Philosophy of Science: Suppose theories are underdetermined by observations or evidence. How then do scientists decide between two competing hypotheses that are both compatible with the data?

II. ULTIMATE VS. NON-ULTIMATE DESIRES

> Purely Instrumental Desire (S&W def.): S wants m ***solely as a means*** to acquiring e if and only if S wants m, S wants e, and S wants m *only* because she believes that obtaining m will help her obtain e. (I.e. S wouldn't want m if she ceased to believe it would help her obtain e.)
>
> Non-instrumental or "ultimate" desire: S wants x as ***an end in itself*** if and only if S wants x and there is no y≠x such that: (1) S wants y, (2) S believes that obtaining x will help her get y, and (3) S wouldn't want x if she didn't want y and believe that obtaining x would help her get y.

Observe that an agent can want something both as an end in itself and as a means (though not "solely" as a means). In such a case our definition says that the agent's desire is non-instrumental or ultimate. Let's call desires in this category **hybrid desires**. Question: If all our non-instrumental desires for the good of others were hybrid desires such that we wanted the good of another person both for its own sake and because it would promote our own wellbeing, could we be altruists?

Plausible examples of non-instrumental desires: your desire to eliminate your own physical pain and your desire to promote your own physical pleasure.

Question: Are there non-instrumental desires other than the desire to avoid experiencing physical pain and the desire to experience physical pleasure?

III. PHILOSOPHICAL ARGUMENTS AGAINST HEDONISM

Psychological Hedonism (S&W def.): All of our *ultimate* motives are desires for our own pleasure (i.e. positive experiences) or aversions to our own pain (i.e. negative experience) . Whenever we want some other person to do well (or badly) we have that desire *only* because: (a) we have a self-directed desire for some pleasure x, and (b) we believe that if that other person does well (or badly) we will get x (or are more likely to do so).

Task: Describe Nozick's experience machine. Would you hook yourself up to the machine? Why or why not?

How can the Hedonist explain why many of us would not go into the machine? Sober and Wilson's answer, "Hedonism is not betraying its own principles when it claims that many people would feel great contempt for the idea of plugging in and would regard the temptation to do so as loathsome. People who decline the chance to plug in are repelled by the idea of narcissistic escape and find pleasure in the idea of choosing a real life."

So suppose I decide not to go into the machine. S&W's idea is that my actions are guided not by beliefs about the experience machine and whether it would bring more pleasure than would regular life, but rather my beliefs about my own states of consciousness. (They can't be saying that my actions are guided by those states of consciousness themselves—i.e. the pleasure and pain of the thoughts—because this would support a non-belief-relative form of hedonism—which is demonstrably false.) According to the proposal S&W think is plausible: I want to maximize my pleasure and I believe that "the idea" of plugging in is painful ("contemptible") and the temptation to do so is painful (i.e. loathsome) and I believe that deciding not to plug in would be pleasurable, so I decide to make the decision not to plug in and to rid myself or avoid the painful idea and temptation to do otherwise.

Question: How plausible is this hypothesis? Shouldn't I realize that I can equally rid myself of these contemptible thoughts by plugging into the machine? Does introspection place any constraints on psychological theories? What sort of psychological reality are these beliefs about my own experience supposed to have? Can't subjects act prior to having such beliefs? Is the egoist S&W describe committed to saying that we are irrational to decide not to go into the experience machine? Must they say that the decision is imprudent (i.e. motivated by a desire for present pleasure of degree n, that is motivationally stronger than our desire for a much greater pleasure n+m)?

IV. THE PROPOSITIONAL ATTITUDE ANALYSIS OF DESIRE

S&W claim that all desires have *propositional content*. Whenever you want something, or want something to occur, this is best represented as your **wanting that s**. Here 's' is to be replaced by a sentence, and 'that s' is to be understood as denoting a proposition: something that can be true or false. So if you want a ham sandwich, this is best represented as your wanting that you have a ham sandwich, or your wanting that you eat a ham sandwich, or your wanting that you get and eat a ham sandwich.

> Question: How plausible is the claim that all desires or motivational states are propositional? How about the desires of a non-human animal or infant? Do non-human animals and infants have desires? Are they "conceptually articulated"? Can someone have a desire with a given propositional content even though she lacks the concepts we use when specifying that content? For example, can a dog or infant desire his food or milk (or desire that he have food or milk) without having a concept of himself?

The Distinction between Cause and Content

By casting things in this way, S&W can allow for a conceptual distinction between the *causes* of a desire and its *content*. The two needn't be the same. So, for example, if a hypnotist conditions me to want a ham sandwich whenever the bell strikes ten, that doesn't mean that I have a desire for the hypnotist or a desire for the way the hypnotist made me feel; if a doctor sends electrical impulses into my brain and this makes me want a ham sandwich this needn't imply that I want the electrical impulse or want the doctor to do what he is doing. Similarly, if I want that I eat a ham sandwich because I gained gustatory pleasure from eating such things in the past this doesn't *immediately* imply that I now want the gustatory pleasure. Suppose my belly is full, and I know that I've scorched my taste buds and lost my sense of smell to a horrible cold, but that I nevertheless continue to want a ham sandwich. Here I have a gluttonous desire for a ham sandwich even though I know it will not give me pleasure despite the fact that my getting pleasure from ham sandwiches in the past causally explains why I now want a ham sandwich.

On the Distinction Between Proximate Motives and Mechanisms of Selection:

Evolution via individual selection: When a phenotypic trait increases the reproductive fitness of the organism with that trait in comparison to other members of its population and thisdifference in fecundity explains why the trait is possessed by greater numbers of individuals in that population over time, the trait has evolved via individual selection. (E.g. white fur for artic foxes.)

Evolution via kin selection: When a phenotypic trait increases the reproductive fitness of an organism's kin group in comparison to other kin groups in a population, and this explains why the trait is possessed by greater numbers of individuals in that population over time, the trait has evolved via kin selection. (E.g. the sterility of "worker" bees.)

A trait can increase both individual fitness and the fitness of kin. In such a case it can evolve via both individual and kin selection. But "evolutionarily altruistic" traits are defined as those that diminish individual fitness. (The sterility of worker bees is an example.) If these traits are adaptations, they must have evolved via kin selection or some other form of group selection as Darwin and S&W both argue.

Notice the difference between "kin selection" and "preference for kin": a phenotypic trait (such as sterility in bees) may evolve via kin selection and yet have nothing to do with preference for kind (as sterility does not involve a preference at all and so does not involve a preference for kin).

Drawing the distinction between the cause and content of a desire allows us to see that **altruism as Sober and Wilson describe it can evolve via individual selection or kin selection and still be altruism**. Suppose Jill gives Elaine her food to relieve Elaine's suffering and that this is motivated by Jill's ultimate desire for Elaine's wellbeing or happiness. Jill's act is altruistic according to Sober and Wilson's definition of altruism even if Jill's desire that Elaine experience happiness is caused by hypnosis or a microchip implanted by a meddling scientist. (The causes of Jill's desire don't matter for altruism, only the content of the motivating desire and whether or not it is ultimate.). And if an altruistic desire can be caused by hypnosis it can also be innate, where the distal explanation for this component of a person's innate psychology is kin selection or individual selection. According to S&W, it doesn't matter whether an ultimate desire to help others evolved via natural selection and is genetically inherited or whether an ultimate desire to help others inevitably has a source in cultural learning (e.g. religious instruction). Again, the causes of the desire are not relevant to the question of whether it generates genuinely altruiostic actions.

<u>Question</u>: Is this a problem with S&W's definition of altruism? Do the causes of the purported ultimate desires for the wellbeing of others matter to us when we assess the action and judge whether or not it was altruistic in nature?

V. SELF-DIRECTED VS. OTHER-DIRECTED DESIRES

> <u>Self-directed desires (Primary Def)</u>: X's desire D is self-directed if and only if D's content involves X or a first-person representation of X, and does not involve some agent Y≠X (or a representation of an agent Y≠X).
>
> <u>Other-directed desire</u>: Desires that are not self-directed.

<u>Questions</u>: Suppose that I want that Mary kisses me and this isn't grounded in another desire. Is this desire self-directed or other-directed? S&W say that Psychological Egoism is distinct from Hedonism in that of the pair only Hedonism is committed to Motivational Solipsism. But this seems wrong given their definition of 'self-directed desire'.

They do discuss **relational desires** (like my desire that Mary kiss me), which have people or representations of people other than the agent herself (or representations of people other than the agent herself) in their contents along with the agent herself (or a first-person representation of such). But they say that the claim that we have some ***ultimate*** *relational desires* is distinct from both psychological egoism and psychological altruism. But this is wrong given the way they've defined psychological altruism as the mere denial of psychological egoism.

> <u>Psychological Altruism</u> (Primary Def.): There is some agent A who Xs because she wants that p, where: (a) A's desire that p is ultimate (i.e. not *entirely* instrumental), and (b) the content of A's desire involves benefitting, helping or aiding some agent other than A (or a representation of an agent other than A).

If psychological altruism is defined in this way, I am a psychological altruist if I have an ultimate desire that Mary return my love, and psychological altruism exists so long as motivational solipsism is false.

S&W could define 'self-directed desire' differently.

> <u>Self-directed desires (Alt Def)</u>: X's desire D is self-directed if and only if D's content involves a benefit or something positive for X (or involves the idea of benefits accruing to X under a first-person representation of X as the agent in question).

The claim that some ultimate desires for a benefit are not self-directed in the sense provided by this alternative definition yields a different division between psychological egoism and psychological altruism. Here the existence of psychological altruism contradicts **psychological relationalism.**

> Psychological Altruism (Alt Def.): There is some agent A who Xs because she wants that p, where: (a) A's desire that p is ultimate (i.e. not *entirely* instrumental), and (b) the content of A's desire concerns benefiting someone B≠A, and does not involve A (or a first-person representation of A) at all.

So if there is one agent who sends a check to the UN mission in Monrovia because she has an ultimate desire that the Liberians enjoy peace, then psychological altruism is true. But if this agent sends the check because she wants that *she* help the Liberians enjoy peace, psychological altruism is not therein made true. Is this a correct way to think of altruism?

VI. PSYCHOLOGICAL EGOISM AND THE HUMEAN THEORY OF MOTIVATION

> **S&W point out that Psychological Altruism "so defined" is not ruled out by the fact (if it is a fact) that every action is the product of a desire.** Why would someone think that unless we act from something other than our desires, we never act altruistically?

Hobbesian theory of desire: Every ultimate desire is for the pleasure or wellbeing of the person who experiences that desire.

Belief/desire thesis: Every action is motivated by one or more of the agent's desires and her belief that performing the action in question will satisfy those desires.

(1. Hobbesian theory of Desire + 2. Belief/Desire thesis) → 3. Impossibility of altruism

Kantian requirement on altruistic action: To act altruistically, an agent must act from her sense of duty or her knowledge of what she is obligated to do independently of any desires or inclinations she might have to perform the action in question.

The Kantian Requirement on Altruistic Action might arise from acceptance of the Hobbesian theory of desire. If all of our ultimate desires are self-directed, we would need to act from something other than desire to act altruistically, which would require the abandonment of the belief/desire thesis.

Humean theory of desire: Observation reveals that humans have a number of different ultimate desires. We desire our own pleasure and freedom from pain, but we also commonly have ultimate desires to the happiness of our friends and family members and for the suffering of our enemies.

¬((1. Humean theory of Desire + 2. Belief/Desire thesis) → (Impossibility of Altruism))

Assignment: Explain or interpret the negated conditional above.

VII. DIFFERENT VARIETIES OF ALTRUISM

E-Over-A Pluralist: Prefer the other benefit rather not in absence of conflict; prefer benefit for self rather than not; prefer benefit for self to benefit for other when they conflict.

Pure Altruism: Prefer the other benefit; indifferent toward benefit to self.

A-Over-E-Pluralist: Prefer the other benefit rather not in absence of conflict; prefer benefit for self rather than not; prefer benefit for other to benefit for self when they conflict.

S&W: Even E-Over-A pluralism is a form of altruism if preference for the other person's welfare is ultimate.

Questions: Is this right? Do amounts or degrees of preference matter here? Suppose I'm allowed to choose between the following three offers: (a) I get $99 and you get $0; (b) I get $98 and you get $50 and (c) I get $99 and you get $25. If I choose (c), is that act made altruistic by my preference for (c) over (a)? Mustn't I forgo $1 to get you $25?

VIII. ALTRUISTIC DESIRE v MORAL PRINCIPLE

1. Principles must be general; altruistic desires can be irreducibly particular.

2.S&W seemingly suggest that radical deontologists can act on moral principles for no other reason than their belief that they ought. What is perhaps more obvious is that people can act from moral principles out of fear of divine punishment, but that they fail to act altruistically in such cases.

IX. BATSON'S EXPERIMENTAL PROOF OF ALTRUISM

Empathy v. Personal Distress: The former decreases heart rate; the later increases it.

S&W on Empathy: S empathizes with O's experience of emotion E if and only if O feels E, S believes O feels E, and this causes S to feel E for O.

What is it to feel E for O?

"If S feels sad for O, then S forms some belief about O's situation and feels sad that this proposition [i.e. the proposition she believes] is true. When Barbara empathizes with Bob, he is the focus of her emotion; she doesn't just feel the same emotion that Bob experiences. Rather Barbara feels sad that Bob's father has just died; she feels sad about what has made Bob sad" (p. 234).

Question: Can you empathize with someone even if you don't feel what they feel? What are the respective roles played by: (a) interpersonal understanding (i.e. thinking of things from another person's point of view to gain knowledge of their thoughts and feelings), and (b) emotional response (i.e. responding appropriately to what you know the other person is thinking or feeling)? Can you give a better characterization of empathy?

> Initial questions: Describe Batson's experimental setup. How does Batson propose to instill empathy in some subjects and not others?

Batson's Two Theses: (1) Empathy (so instilled) causes people to help those with whom they empathize (at a cost to themselves). (2) Empathy causes empathetic people to help by inducing in them an ultimate (or non-instrumental) desire for the wellbeing of those with whom they empathize.

The empathy-altruism hypothesis: Empathy causes people to help those with whom they empathize from altruistic ultimate desires: i.e. ultimate desires for the wellbeing of the parties in question.

In particular, the role empathy plays in causing people to help others cannot be attributed instead to desires posited by the following theories:

(a) **The aversive arousal hypothesis**: empathetic subjects desire to rid themselves of the experience of watching people suffer (as this is more uncomfortable for them than it is for other subjects).

(b) Empathetic subjects have a greater aversion to painful memories of the suffering person or fear these memories more than low empathy subjects,

(c) **Empathy-specific punishment hypothesis 1**: Empathetic subjects have a greater desire than low empathy subjects to avoid the negative feelings they think they would get if they didn't help and other people censured them for their failure to help,

(d) **Empathy-specific punishment hypothesis 2**: Empathetic subjects have a greater susceptibility to guilt or a greater desire to avoid the negative feelings of guilt they think they would get if they didn't help than do low empathy subjects.

(e) **The Empathy-specific reward hypothesis**: Empathetic subjects desire to experience some kind of mood-enhancing feeling they think they will get from realizing they have helped another person or more highly anticipate this kind of reward than do low empathy subjects.

(f) **The Empathetic joy-hypothesis:** Empathetic subjects desire to experience some kind of mood enhancing feeling they think they will get from realizing the person has been helped (whether by them or by someone else) or more highly anticipate this kind of reward than do low empathy subjects.

(g) Empathetic subjects desire to rid themselves of the sadness they feel when they experience empathy upon perceiving the plight of another, which they experience more acutely than do low empathy subjects.

Batson proves (1) by showing that people are more likely to help another person in need if they empathize with that person (so long as empathy is reliably augmented in the ways he imagines). The argument for (2) involves (a)-(g). For each of (a)-(g) Batson considers the hypothesis that the helping behavior he has isolated is explained by that motive alone. He then contrives further experiments to test whether each one of these motives is present in those individuals whose empathy leads them to help. In each case he concludes that the egoistic hypothesis in question makes predictions about how empathetic people will behave that are not born out by the data.

In regard to (a): The aversive arousal hypothesis predicts that empathetic subjects are less likely to help when it's easier to leave the scene. But they aren't.

In regard to (b): Hypothesis (b) predicts that empathetic subjects will be more likely to choose to receive news about the situation of the person in need if they think this news will be good. But they aren't.

In regard to (c) and (d): The empathy-specific punishment hypotheses predict that empathetic subjects are less likely to help when it would be easier for them to justify their not helping (the ease in justification coming from the information that many others in their situation had declined to help). But they aren't.

In regard to (e): The empathy-specific reward hypothesis predicts that empathetic subjects are less likely to help when they know (or believe) they can get a mood enhancing experience without helping. But they aren't.

In regard to (f): the Empathetic-Joy hypothesis predicts that empathetic subjects will be happier if they believe that they have helped the other person than they will be if they believe the other person has been helped by people other than themselves. But they aren't.

In regard to (g): Hypothesis (g) predicts that empathetic subjects are less likely to help when they know (or believe) they can get a sadness-mitigating experience without helping. But they aren't.

(Hypothesis (g) also predicts that empathetic subjects should be less likely to help if they think their current mood is frozen in place. Empathetic subjects were led to believe this—by being told they had been given a mood freezing drug—but the experimental results concerning whether this detracted from the likelihood of their helping was equivocal.)

S&W seem to allow that Batson is right about the experimental predictions each theory makes, and he's right to conclude that the data show these predictions to by contradicted by the experimental data. Still, they insist, Batson's experiments don't rule out the hypothesis that empathy causes people to help others by creating a desire to get rid of a particular kind of sadness that (subjects believe) can only be assuaged by helping or hypothesis (d) that empathetic subjects act to avoid a particular feeling of this kind—a feeling with which S&W equate guilt.

Question: How plausible is the hypothesis that there is a desire to avoid such a feeling of guilt involved in all acts of helping? (Does our reaction to Nozick's experience machine contradict the hypothesis that our acts of helping are motivated by a purely non-instrumental desire to avoid or rid ourselves of such a feeling?) You will choose a massage over a handshake because you rightly believe the former to be more pleasurable than the latter. Is the guilt always more painful or bad for you than the costs of helping will be? If not, how can the altruism-denier argue that you always mistakenly believe that it will be when you help another knowing it will cost you?

Question 2: Given their reactions to Batson's results, why don't Sober and Wilson accept the conclusion they attribute to Wallach and Wallach (1991) on which no experimental refutation of egoism is possible?

The Importance of the question: People who believe in egoism are less likely to help others. See the experiments on economics students conducted by Frank, Gilovich and Regan (1993); discussed on pp.273-4 of S&W.