Regretful Decisions and Climate Change

Rebecca Livernois

Climate change has made pressing the question of why we do little to significantly reduce greenhouse gas emissions. Chrisoula Andreou (2006) argues that even when interpersonal conflicts of interest have been resolved, we may continue to destroy the environment because environmental problems have a similar structure to the puzzle of the self-torturer (PST). Therefore, she claims, environmental problems have a similar outcome to the PST: coherent yet intransitive preferences. I argue that intransitive preferences are not coherent in the PST and in PST-like environmental problems. This becomes clear when the level of certainty possessed by the self-torturer is taken into account. I suggest instead that aiming for 'optimality' in the sense of balancing the costs and benefits of an activity will likely lead to regret in scenarios, like climate change, that are characterized by uncertainty and individually negligible effects.

Much attention has been given to our seemingly strange preferences for environmental quality. Why is it that we willingly follow paths of environmental destruction when environmental quality is necessary to sustain a high quality of life? Climate change has made this question even more pressing—experts advise immediate action yet we do little to significantly reduce greenhouse gas emissions. Analysis of our willingness to damage the environment has led to two prominent theories. The first theory holds that at least some environmental problems have the structure of a multi-player prisoners' game: if each individual acts in accordance with her private interest in using a common resource (like using the atmosphere as a dump for carbon dioxide (CO_2) emissions) then they all do worse than if they acted in accordance with the interest of the group. The second theory holds that some environmental problems are aggravated by cost distribution issues: the cost of environmental damage is not incurred by those who create the damage, so they have little private incentive to mitigate the damage. Both of these theories are based on interpersonal or intergenerational conflicts of interests regarding the use of common environmental resources. Environmental economic models suggest ways to solve these problems. For example, a tax on CO_2 emissions is intended to relocate the costs associated with pollution from external parties to those who create the damage.

Chrisoula Andreou (2006) suggests that these theories explain some aspects of environmental problems but are not exhaustive. She argues that even when interpersonal conflicts of interest have

been resolved, we may still continue to destroy the environment (Andreou 2006, 99). By analogy to the puzzle of the self-torturer (described below), she argues that in scenarios where there are individually negligible costs and large benefits of an action, a unified collective can follow a path of destruction by making choices based on informed, understandable, yet intransitive preferences (Andreou 2006, 96). Preferences over options A, B and C are transitive when A is preferred to B, B is preferred to C and A is preferred to C ($A \succ B, B \succ C, A \succ C$). The type of intransitive preference in question here those that form a preference loop: A is preferred to B, B is preferred to C, and C is preferred to A ($A \succ B, B \succ C, C \succ A$). Andreou's claim is more significant than it may initially seem: not only does she suggest the existence of coherent intransitive preferences, but furthermore suggests that some environmental preferences in particular tend to be coherent and intransitive. If environmental preferences were intransitive, then many environmental economic models would be useless at providing solutions to environmental problems because these models require the assumption of transitive preferences. In particular, any model that includes consumers, like those that estimate a policy such as an optimal tax on carbon emissions, would be useless if climate change preferences were characteristically intransitive. Economic theory is a theory of averages, so it is unproblematic if some people have intransitive preferences some of the time as long as, on average, people have transitive preferences. However, if people have intransitive preferences on average—if environmental preferences are characteristically intransitive—then this theory would no longer be accurate. It would be incapable of appropriately modeling decision-making behaviour because its foundational assumptions would be essentially mistaken. As such, my task herein is to determine if it is indeed believable that environmental preferences are characteristically intransitive. I focus on climate change preferences because climate change is arguably the most important environmental problem.

In this paper, I first explain the puzzle of the self-torturer (PST) and Andreou's argument that environmental problems have a comparable structure and therefore comparable result to the puzzle. I then argue that Andreou is right to claim that environmental problems have a similar structure, but reject that intransitive preferences are coherent in the PST. This becomes clear when the level of certainty possessed by the self-torturer is taken into account. That is, I argue that in PST-like scenarios that are characterized by uncertainty, it can appear as though the self-torturer exhibits intransitive preferences when really she just exhibits regret. This claim rests on my argument that if

the self-torturer possesses certainty, then intransitive preferences are incoherent.[1] I conclude that climate change decisions are made under conditions analogous to the PST that includes an element of uncertainty. In this type of scenario, a unified collective can end up in an understandable, yet regrettable situation by acting on transitive preferences. This is significant because it avoids the troubling implications of the existence of understandable, or coherent, intransitive preference and offers an alternative explanation for why a unified collective that makes decisions about climate change can understandably end up in a regrettable state.

1. The Puzzle of the Self-Torturer

Consider Warren Quinn's (1990) puzzle of the self-torturer (PST).

Suppose a patient, called the self-torturer, wears a portable electric shocking device. This device has a dial of ordered electricity-level settings from 0 (off) to 1001. Increased settings correspond with increased pain—at level 1 there is no pain while at 1001 the pain is excruciating. For each single turn of the dial the electricity level increases so slightly that the self-torturer cannot discriminate between adjacent settings. However, she experiences different pain levels at sufficiently distant settings. Before the experiment begins, the self-torturer has a trial week to test and compare different settings. At the end of the week, the dial is returned to 0. From then on, she has two options each week:

A. Stay at the current setting and receive no reward,
B. Increase to the next setting and receive $10,000.

If the self-torturer chooses option (A), the experiment ends and she indefinitely stays at the current setting. If the self-torturer chooses option (B), then she is presented with the same two options the next week. The outcomes of the options are the same each week; she never has the option to return to a lower pain level. She is aware of all these conditions. The following assumptions are made about her preferences:

i. The self-torturer prefers more money to less money and less pain to more pain,
ii. There is some setting s_m such that for any setting s_n where $n \geq m$, the self-torturer prefers setting 0 to remaining at s_n and getting the associated reward,

[1] I use 'certainty' to mean that the agent possesses full information and perfect foresight in the context of the experiment. This means that she knows *ex ante* everything she would know *ex post* for every possible choice.

iii. No other preferences are relevant for the self-torturer's choices.[2]

The puzzle arises from considering what would be reasonable for the self-torturer to do under these conditions. Given preference (i), the self-torturer should invariably choose option (B) each week because given that adjacent settings feel the same, the only relevant preference each time she chooses between the two options is her preference for more money. If she chooses (B) each week then she will end up at the last setting where she is in excruciating pain. If her preferences were transitive, then she would prefer setting 1001 to setting 0 because she prefers increasing the setting each week ($s_{j+1} \succ s_j$). However, understandably, she does not prefer setting 1001 to setting 0 because she prefers no money to being tortured, in line with preference (ii). This is taken to show that the self-torturer has intransitive preferences: s_{m-1} is less preferred to s_m which is less preferred to s_n which is less preferred to a lower setting such as s_{m-1}. Moreover, once she returns to s_{m-1}, she has reason to continue choosing option (B) leading her back to s_n; that is, her preferences are cyclical ($0 \prec 1 \prec \ldots \prec s_{m-1} \prec s_m \prec s_n \prec 0$). Therefore, the self-torturer is in a tricky situation because there seems to be no optimal stopping point. The self-torturer seems to be right to think that if she is going to quit, she is always better off quitting at the next setting rather than the current setting. Yet she is also right to think that if she always acts this way she will end up in excruciating pain. The self-torturer's problem is that any stopping point is suboptimal; even stopping at a point before s_n is suboptimal because she could have been better off stopping at the next setting, since it feels identical.

The assumptions about the self-torturer's preferences seem ordinary because it is normal to prefer more money and less pain and to prefer poverty to being tortured indefinitely. The result is that seemingly ordinary preferences can lead to a violation of transitivity, which is a central constraint on rational choice. That is, contrary to the tradition in rational choice theory, the puzzle is supposed to show that having intransitive preferences is sometimes coherent. This therefore presents a challenge to the widely held idea that transitivity is a prerequisite to rationality.[3]

[2] This set-up is based on the way Sergio Tenenbaum and Diana Raffman (2012, 88-9) structure the PST.

[3] One argument for this requirement is that an agent with intransitive preferences can be made into a money pump, which is not rational behaviour. Suppose that an agent prefers cake to steak, steak to salad, and salad to cake. She is willing to pay $1 to have her preferred option. I can give her salad and then offer her steak for $1, which she would accept. I then offer her cake for $1, which she would accept. I then offer her salad for $1, which she would also accept. She thus ends up where she started – with salad – yet she is $3 poorer.

The Puzzle of Air Pollution

Andreou constructs a simplified case of an environmental problem with a structure similar to the PST. In this puzzle the decisions-maker is a unified collective, which rules out interpersonal conflicts of interest. Each month it decides whether to continue consuming pollution-intensive luxury goods. The pollution is known to be carcinogenic but its effect on health each month—the marginal cost of consumption—is negligible. The unified collective knows that if it does not reduce pollution eventually, the health of the community will be seriously damaged (Andreou 2006, 104). It has the following choices each month:

A.e. Stop consuming luxury goods and thereby reduce pollution,

B.e. Continue consuming luxury goods and thereby continue polluting.

It has the following preferences:

i.e. The unified collective prefers more consumption of luxuries to less consumption of luxuries, and less pollution (more health) to more pollution (less health),

ii.e. There is some level of health s_m such that for any level of health s_n where $n \geq m$, the unified collective prefers setting 0 to remaining at s_n and getting the associated reward,

iii.e. No other preferences are relevant for the unified collective's choices.

This scenario is structurally identical to the PST where pain is replaced with poor health, monetary reward is replaced with consuming luxury goods, and increasing voltage settings is replaced with monthly pollution. The same tricky situation arises for the unified collective: since one extra month of pollution makes no noticeable difference to the community's health, the only preference that is relevant each month is the preference for consuming more luxury goods. Therefore, it is always better for the unified collective to keep consuming luxuries for one more month. However, this path leads to a state where health is so poor that it would give up all the rewards to return to a better state of health. Hence, the unified collective's preferences are intransitive (Andreou 2006, 105).

I interpret Andreou's argument as follows:

1. If key PST conditions arise, then intransitive preferences arise,

2. These key PST conditions arise in some environmental problems such as air pollution,
3. Therefore intransitive preferences arise in some environmental problems such as air pollution.

I also interpret Andreou to take individually negligible effects as the key PST condition. She states: "where individually negligible effects prompt intransitive preferences, destructive conduct can prevail even if the (individual or collective) agent is guided by a single (or shared) set of stable and informed preferences" (Andreou 2006, 104-5). As such, she interprets the PST to show that individually negligible effects lead to intransitive preferences. Since environmental problems also have individually negligible effects, a unified collective has intransitive environmental preferences. In the next section I argue that premise 1 only appears to be correct when the level of certainty possessed by the self-torturer is not taken into account.

3. Objection

Premise 2 of Andreou's argument seems accurate when we suppose, along with Andreou, that individually negligible effects is the key PST condition. That is, it is believable that the effects of climate change are individually negligible because the effects are noticeable only at large time intervals, such as every decade (Brown 2014, 131). Notice, however, that the PST does not require that individually negligible effects are imperceptible in order to produce the result of sensible intransitive preferences (Tenenbaum and Raffman 2012, 11). Even if marginal costs are perceptible yet minuscule in comparison to the marginal benefits, the self-torturer still has an incentive to continue to the next setting each time she contemplates stopping the experiment because the benefits far outweigh the costs of continuing. Therefore, even if it is not believable that the costs of climate change are not imperceptible each time we are faced with a climate change decision, it is conceivable that the perceived short term costs associated with an extra period of polluting are dwarfed by the benefits we receive from polluting activities. As such, the marginal costs of climate change are analogous to the negligible marginal costs in the PST.

I argue that premise 1, on the other hand, is mistaken. As it has thus far been described, it is unclear what role certainty plays in the PST. It is important to explicitly consider the role played by certainty in the two puzzles for two reasons. The first reason is that assumptions about the amount of information available to the self-torturer and the unified collective are required to maintain the

analogy between the PST and the puzzle of air pollution. Recall that the self-torturer has a trial week at the beginning of the experiment to discern the relationship between her pain and the settings. This gives her knowledge about the facts of the experiment before she begins making decisions about whether to increase or to remain at the current setting. This level of information is not available in environmental problems in reality. Indeed, a key difficulty of environmental decision-making is that outcomes are inherently uncertain (Broome 2012, 117). For example, it is uncertain how many months of polluting activities results in an unbearable level of pain. If the unified collective was unaware of the precise health impacts from the amount of time spent polluting, it would be believable that its ignorance, and not necessarily intransitive preferences, led it into a terrible state: perhaps it did not curtail its pollution earlier because it did not realize that it was causing such harm. Therefore to maintain the analogy between the two puzzles, we must assume that the unified collective knows that an accumulation of pollutants will inevitably lead to poor health in a certain number of months, just like the self-torturer knows that the pain is unbearable, but capped, at setting 1001. We must therefore assume that the unified collective has much more information than it would have in reality about the outcomes of its polluting behaviour on the environment and health.

The possibility of ignorance resulting in a regrettable outcome points to the second reason for considering the level of certainty in the PST: intransitivity can only be exhibited unambiguously when the decision-maker acts under certainty. In the PST, we 'observe' what we take to be reasonable decisions. Decisions are assumed to be revealing of preferences, which are not directly observable themselves. That is, it is assumed that reasonable agents choose their most preferred option. However, uncertainty can disrupt this direct link between preferences and decisions. When an agent is uncertain, her choices are unlikely to be solely informed by her preferences. Instead, it is likely that her choices are based on her preferences in addition to her expectations of future outcomes. The following discussion of premise (c) expands on this notion. Therefore, to interpret decisions as unambiguously revealing of preferences, we must observe decisions that are made under certainty.

The self-torturer's preferences are supposed to be a cyclical form of intransitive preferences: A is preferred to B, which is preferred to C, which is preferred to A. If this is the self-torturer's informed preference structure and her preferences are stable, then no amount of additional knowledge would change her preferences. That is, the self-torturer is supposed to exhibit a stable preference loop:

having reached an excruciating setting, she prefers to return to a lower setting; once she is there, she prefers to increase the settings until she is back at the excruciating setting, *ad infinitum*. So, to determine if the self-torturer's behaviour is really the result of intransitive preferences, then we must observe this cyclical behaviour under certainty. If she was acting under uncertainty and had transitive preferences, we could observe something that looks like intransitive preferences—due to ignorance she reaches a state that she regrets. However, as she gained knowledge, she would eventually learn how to avoid a regretful scenario. As such, her decisions would not be cyclical. Therefore, to determine if the behaviour of a decision-maker is the result of intransitive preferences, we must eliminate the confounding effects of uncertainty on her behaviour. In any case, the PST is meant to show that even assuming away other factors that make good decision-making difficult over time—like uncertainty—the self-torturer still ends up in a terrible state because of her intransitive preferences. Therefore, I contend that it is important to analyze the reasonable behaviour of the self-torturer in an idealized case where she possesses certainty, which involves possessing full information and perfect foresight.

In sum, I rely on the notion that having intransitive preferences in normal circumstances is incoherent, even if it occurs in practice. The PST is taken to show that intransitive preferences are coherent in certain circumstances, meaning that we can understand why one would have such preferences. In the next section I argue that in PST-like circumstances, like in normal circumstances, intransitive preferences are incoherent. This becomes clear when we assume the self-torturer possesses certainty. In the subsequent section, I argue that some level of uncertainty is required to generate the result of the self-torturer understandably ending up in a regretful state. That is, I suggest that the PST only appears to reveal the existence of coherent yet intransitive preferences when the level of certainty in the experiment is not explicitly considered. Therefore, I suggest that the level of certainty plays a large, but unacknowledged role, in the PST. My argument proceeds as follows:

a. In the PST, either the self-torturer possesses certainty or she acts under uncertainty,
b. If the PST conditions include certainty and negligible marginal costs, then intransitive preferences are not sensible,
c. If the PST conditions include uncertainty and negligible marginal costs, then intransitive preferences are not sensible,
d. Therefore, premise (1) is incorrect.

Consider premise (b)

Consider the case where the self-torturer makes decisions under certainty: she possesses full information and perfect foresight. This is not a stretch from the original specification of the PST because the PST sets out conditions that imply that the self-torturer is well-informed: she has a trial period where she can test different settings, she does not suffer from cognitive failures, her preferences are stable and she is aware of her preferences. I add to this that the self-torturer has all knowledge she may require during the experiment.

Recall preference (ii) which states that there is some setting s_m such that for any setting s_n where $n \geq m$, the self-torturer prefers setting 0 over remaining at s_n and receiving the associated reward. The assumption of certainty means that the self-torturer can always locate s_m; however, if the self-torturer can locate s_m then it is never reasonable for her to proceed past s_m. This is because the self-torturer gets \$10,000 for moving from s_m to s_n, but at s_n and not at s_m she is willing to give up much more than \$10,000 to return to setting 0. Hence if she were to choose to move from s_m to s_n she would knowingly make herself worse off. Since she is aware of this when she possesses certainty, it is never reasonable for her to reach s_n. So, contrary to the conclusion drawn in the PST, it is never reasonable for the self-torturer to reach a point so painful that she would rather give up all her rewards to return to setting 0.

The self-torturer could reveal intransitive preferences, however, even if she stops before s_n. She could reach a setting where she would prefer to give up some money to return to some lower setting but not necessarily all her money to return to setting 0. If she can determine the level at which she would give up all her rewards to return to level 0, then it is reasonable to suppose that she can also determine other levels at which she would give up none of her rewards and levels at which she would give up some of her rewards to return to some lower setting. If this is the case, then she should be able to locate her tipping point, s_t which is the last setting at which she is not willing to return the most recent monetary reward to return to a lower setting. Hence s_{t+1} is the first level where the self-torturer would be willing to return \$10,000 to return to s_t. This means that the move from s_t to s_{t+1}, if undertaken, is the first point in the experiment where the self-torturer would exhibit intransitive preferences. If the self-torturer can locate s_t because she possesses certainty, however, then it is not reasonable for her to move from s_t to s_{t+1}. Even if she is aware of the fact that this move creates a negligible change in pain level and a large monetary reward, if the self-

torturer is operating under certainty then she knows that this move will bring about just enough additional pain that she is made worse-off by this move, apparent in the regret she knows she will feel once she reaches s_{t+1}.[4]

This may initially sound absurd: how can the self-torturer think she will be worse-off at s_{t+1} than at s_t if the move from s_t to s_{t+1} feels the same to her? And if this is the case, then it seems reasonable, like at any other setting, for her to move from s_t to s_{t+1} because her only relevant preference is her preference for more money. To see why this is not the case, consider a version of the PST where the only difference is that the setting is increased once a week but options (A) and (B) are given to the fully-informed the self-torturer at five-week intervals. The self-torturer cannot feel a difference between adjacent settings but can feel a significant difference in pain between setting-intervals of four. Suppose the self-torturer is in a decision-making week at setting 95 and is trying to decide whether to continue to setting 100 or to stop at 95. While the move to setting 96 creates a negligible increase in pain, the level of pain she knows she will experience at level 100 will be noticeably different to her current pain level. As such, the relevant marginal costs are no longer negligible. This means that she no longer has an incentive to continue to the last setting, since the relevant marginal costs are large enough that she is not always better off if she stops at the next setting. That is, the change in pain level between 90 and 95 is worth the extra $40,000 but she knows that at level 100 the increase in pain is not worth the extra $40,000. She therefore prefers level 95 to level 100. Hence this scenario is not a puzzle; it is a normal situation where a tradeoff is made between the costs and benefits of an action. This means that in this situation, if preferences are informed and understandable, then they are transitive.

If the self-torturer possesses certainty, it is not reasonable for her to act differently in the one-week (original PST) and five-week (altered PST) cases. In the altered PST, it is apparent to the self-torturer that somewhere in the range of 96 to 99 (inclusive) her pain level reaches a point such that

[4] Frank Arntzenius and David McCarthy (1997) make a similar point. They alter the puzzle such that in the self-torturer's trial period before the experiment begins, settings are administered to the self-torturer at random. She reports her level of pain at each setting, which is recorded. The frequency of her pain reports at each setting is given to her and she uses this report during the experiment to differentiate between adjacent settings that feel similar. They argue that the self-torturer should assign utility values to each setting, which decrease as the pain level increases (which is determined by the frequency of reports of pain). She should stop just before the trade-off between monetary reward and pain level switches in favour of the latter. Hence they explain how the self-torturer can achieve full information; I am assuming that the self-torturer has full information and therefore already knows which settings are slightly more painful than others. My solution is the same as Arntzenius and McCarthy: the self-torturer should stop at her tipping point. This is a standard result of optimizing behaviour in rational choice theory—the optimal point is either where marginal costs exactly offset marginal benefits or the point just before marginal costs exceed marginal benefits (a corner solution).

the extra monetary reward is no longer worth the extra pain.[5] The decision is more difficult in the one-week case, however, because the self-torturer must choose which setting to stop at in the range of 96-99 when adjacent settings feel similar. Unlike the altered PST, the point at which her level of pain shifts from 'worthwhile' to 'not-worthwhile' is vague. If the self-torturer stops at 95, she is inclined to move to 96 because it feels similar, and so on until she reaches a level of pain she regrets reaching. This is just the tricky situation described in the original explanation of the PST. However, if the self-torturer can distinguish between the pain level at setting 95 and 100, then there must be some pair of settings between 95 and 100 where, if asked what her pain level is at each setting, her answer shifts from 'worthwhile' to 'not worthwhile'.[6] The last point at which the self-torturer reports that the pain is worth the money is her tipping point. Since the self-torturer is operating under certainty, she can locate this point. Therefore, there is an optimal stopping point; this is the point at which the extra pain from increasing settings is just offset by the extra monetary rewards.

If certainty is explicitly considered in the PST, then the PST conditions do not imply understandable yet intransitive preferences. It is not reasonable for the self-torturer to proceed past s_t because she possesses the knowledge that this would make her worse off, and it is not reasonable to knowingly make oneself worse off. Therefore it is incoherent for the self-torturer to have intransitive preferences in the PST. Therefore premise (1) is incorrect when the key PST conditions include individually negligible effects and certainty.

Consider premise (c)

Given that intransitive preferences are not reasonable when the self-torturer possesses certainty, if there is a puzzle at all then it must be the case that (at least the appearance of) understandable intransitive preferences arise when the key PST conditions include individually negligible effects and uncertainty. I consider two forms of uncertainty that could arise in the context of the PST. Recall that s_t is the tipping point: it is the last setting at which the self-torturer does not prefer to return to a previous, lower setting by paying back some amount of money. Note that each setting (s) is associated with a pain level (p). The relationship between p and s is variable. For example, setting

[5] This could be explained by arguing that the self-torturer has diminishing marginal value of money and increasing marginal disutility of pain. This is a troublesome assumption because these concepts require transitive preferences, which is just the aspect of rationality the PST is challenging. To avoid this, I opt for explaining this just in terms of costs, benefits and thresholds—the self-torturer reaches a level of pain somewhere between 96 and 99 where her long-term interests dictate that she should not increase the pain level any further, because no amount of money can compensate her for being in a certain level of pain for a long period of time.

[6] This is informed by an argument made by Shelly Kagan (2011, 132-3).

500 in one experiment and setting 50 in another experiment could be associated with the same pain level (that is, the amount of electricity associated with setting 500 and 50 in the two experiments is the same).

The first type of uncertainty is uncertainty about the pain level that marks the tipping point, p_t. Like in the original PST, the self-torturer knows the number of settings in the experiment and the pain level at the beginning and end of the experiment. As such she can back out the relationship between p and s; for example, she knows that she feels a significant increase in pain at setting-intervals of four. However, the self-torturer does not know which pain level marks her tipping point and therefore she cannot determine the associated setting at the tipping point. That is, p_t is unknown but the relationship between p and s is known. This could be a result of the self-torturer being unsure how she will weigh the trade-off between pain and monetary reward during the experiment. She could think that she may surprise herself in how little she cares about the money once the experiment starts, or she may think she could become greedier during the course of the experiment.

The second type of uncertainty surrounds the tipping point setting. The self-torturer knows the pain level associated with her tipping point (what p_t is at s_t), but she does not know when, if at all, this setting will be reached during the experiment. This is perhaps because she does not know the number of settings in the experiment, or she does not know the pain level associated with the last setting of the experiment. As such, she cannot back-out the relationship between p and s. Furthermore, she cannot back-out this relationship as the experiment progresses. This could be the case if she thinks that the increase in electricity associated with an increase in the setting is not constant. Sometimes the self-torturer feels the difference after three settings, but other times the increase in electricity for each setting is so minute that she only feels a difference after twenty settings. Hence in the second type of uncertainty p_t is known but the relationship between p and s is unknown, so s_t is unknown.

The first type of uncertainty is ruled out by the assumption that the self-torturer is informed about her preferences and that her preferences are stable over time (Andreou 2006, 105). Furthermore, uncertainty about our surroundings is of interest here, not uncertainty about ourselves, because this is the type of uncertainty we face in climate change decisions. That is, we have an idea of the sort of pain that would make us want to give up the benefits of polluting to return to a higher level of environmental quality (p_t). For example, perhaps there are widespread famines at this level of pain.

However, we do not know precisely when we will reach this level of pain (Broome 2012, 131). It is not uncertain that if polluting habits do not change, then we will reach this painful state eventually, but precise quantitative predictions of the greenhouse effect are uncertain (Broome 2012, 29). So, in climate change decisions we have an idea of what an unacceptable level of pain is (p_t), but we do not know the relationship between years (or amount) of carbon dioxide pollution (s) and our pain level (p) because our pain level is determined by the greenhouse effect. That is, we cannot accurately locate s_t. Since uncertainty is an essential component of climate change decisions, including the second type of uncertainty (which I will now just call 'uncertainty') in the PST (excluding a trial period) is a better model for climate change decisions than the PST and the puzzle of air pollution that assumed a large degree of certainty.

It could be argued that the self-torturer has intransitive preferences when uncertainty is incorporated into the PST. Recall that when the self-torturer possesses certainty, she can locate her tipping point s_t and therefore it is not understandable for her to proceeds past s_t; therefore only transitive preferences are reasonable. However, when the self-torturer makes decisions under uncertainty, she cannot accurately locate s_t and therefore it is conceivable that the negligible marginal costs between adjacent settings interfere with her ability to accurately estimate the setting associated with p_t. Hence she could understandably proceed past s_t and find herself in a scenario in which she would rather give up some monetary reward to return to a lower setting, thus appearing to have intransitive preferences. For example, suppose s_t is setting 96. The self-torturer is at setting 96 but thinks that perhaps setting 97 will feel so similar to 96 that at 97 she will still experience pain level p_t. However, when she takes the gamble and moves to 97 she discovers that she is now at p_{t+1}, which is just enough pain to put her past her tipping point, so she would have preferred to stop at 96. Hence, she seems to have intransitive preferences since she initially prefers 97 to 96, but later prefers 96 to 97.

Intransitive preferences, however, are cyclical when observed over time. The example above only shows that the self-torturer has intransitive preferences if, given the opportunity to go back to setting 96 and resume the experiment, she would again stop regretfully at setting 97. But this would not be reasonable because she would be operating under less uncertainty – she would have a better idea of which setting is associated with p_t and thus (at least given enough tries) should could inform herself of the location of her optimal stopping point, s_t. As such, she would (eventually) behave in the same way as she does in the case of certainty. Therefore, her preferences are not cyclical.

Furthermore, as discussed above, intransitivity can be revealed only under certainty where preference ordering solely determines a decision. That is, under certainty, if S chooses x over y, y over z, and z over x, then the self-torturer reveals that $x \succ y \succ z \succ x$. We can therefore conclude that her preferences are intransitive. However, decisions cannot reveal intransitive preferences under uncertainty even if they have the appearance of intransitivity because under uncertainty decisions are based on expectations. Consider an example of how decisions under uncertainty can have the appearance of decisions that reveal intransitive preferences. Consider a choice between x and z when either x or z must occur, but not both. S must guess whether x or z will obtain and she will receive the associated payoff if she is right and will receive nothing if she is wrong. The payoff of choosing x when x obtains is \$10 and the payoff of choosing z when z obtains is \$1. In the first period, t_1, S deems the probability of x obtaining to be 0.8 and the probability of z to be 0.2. Therefore the expected payoff from choosing x is \$8 and z is \$0.20, so S chooses x over z. In the second time period, t_2, S learns that the probabilities she used to calculate expected payoff in the last period were inaccurate; therefore she updates her probabilities in the next period and subsequently faces the same choice. In t_2, the probability of x obtaining is 0.01 while the probability of z is 0.99. The expected payoff of x is therefore \$0.01 and the expected payoff of z is \$0.99, so S chooses z over x. If it is not taken into account that the decision is based on expectations and not just preference ordering, then S's decisions appear to reveal intransitive preferences since in t_1 she appears to prefer x to z while in t_2 she appears to prefer z to x.

S's decisions, however, are based on her subjective probabilities and her preference for maximizing expected payoffs. So, whether she prefers x or z depends on which she thinks is the most likely to obtain. It is not the case that her preferences are cyclically structure ($x \succ z \succ x$); instead the outcome that maximizes expected payoff is preferred. Even if she chose x in t_1, z in t_2, and x again in t_3, she would still not reveal intransitive preferences because her choices are determined by changes in her subjective probabilities, not necessarily her preference structure. Hence, when we observe decisions that have the appearance of intransitivity but that are made under uncertainty, it cannot be concluded that the decision-maker has intransitive preferences. Suppose that in t_1 S chooses x but z obtains, so she receives zero payoff. She ends up in a regretful state in which she now wishes she had chosen differently in the past, and she chooses differently next time. With hindsight she sees that her probabilities were inaccurate and had they been more accurate, she would have acted like she did in t_2 where she chose z, resulting in a higher payoff. Through experience, she makes a

different, more informed choice next time. It is understandable when a decision-maker who faces uncertainty ends up in a regretful state because outcomes are not always as expected.

This sort of regretful scenario can understandably occur when a unified collective makes climate change decisions. Consider a complicated, and arguably more realistic, version of Andreou's puzzle of air pollution that includes uncertainty. A unified collective makes a decision annually to keep polluting or to stop polluting but it does not know precisely which year is associated with its optimal stopping point. The marginal costs of polluting are individually negligible and the benefits are huge since their economy depends on fossil fuel consumption. There are three expert predictions for how long it will take to reach a certain temperature increase associated with catastrophe (s_m), where each prediction has an associated optimal stopping point (s_t) where the benefits just outweigh the costs of polluting. It believes that p_t will occur in 60 years with a probability of 0.3, in 80 years with a probability of 0.6, and 200 years with probability of 0.1. It thus must decide whether to stop polluting in 60, 80 or 200 years from now. Since year 80 is the most probable optimal stopping point, it plans to stop polluting in 80 years. When it reaches 60 it feels like its pain level is high, but presumes that the next twenty years will feel quite similar to what it is currently experiencing (given uncertainty and individually negligible effects), so it keeps polluting. At year 70, it realizes that it passed its pain level tipping point and now would have preferred to have stopped in year 60. Hence it ends up in a sub-optimal, painful scenario because it expected year 80 to be an optimal stopping point when in fact year 60 was the optimal stopping point. Negligible marginal costs and uncertainty inhibited its ability at year 60 to recognize that it was in fact at its optimal stopping point since it reasoned that an extra 20 years may not feel very different to the pain it was currently experiencing. Hence it ends up in a regretful scenario where, despite choosing to pollute to year 70, it now prefers stopping at year 60 to year 70. This example illustrates how the unified collective can end up in an understandable yet regrettable scenario: despite the unified collective basing its decision on the best information available to it, the chancy nature of the world or its inaccurate probability distributions combined with the difficulty in discriminating between adjacent levels of pain leads it to a point it regrets. This does not show intransitivity, but instead that regretful scenarios understandably ensue under these conditions.

The tricky situation of the self-torturer and the unified collective does not result from sensible intransitive preferences as Andreou suggests; instead I contend that it merely looks like they have reasonable intransitive preferences when it is not acknowledged that they are acting under

uncertainty. The tricky situation results from acting under conditions that make it difficult to locate an optimal stopping point: uncertainty and individually negligible effects make it difficult to determine where the optimal stopping point is because, given the preference for more money and less pain, they always want to push further to maximize benefits, while at the same time it is difficult to locate the point at which the benefits are just offset by the costs because adjacent moves feel similar. That is, individually negligible effects aggravate the difficulty of making decisions under uncertainty because it is easier to end up in a regrettable scenario when it is difficult to determine the actual marginal costs associated with the options, since marginal costs appear to be zero but are invariably nonzero. Therefore we should replace Andreou's first premise with the premise that if key conditions arise in the PST (which include uncertainty and individually negligible effects), then a regrettable scenario can understandably arise. Premise two is unchanged: these key conditions arise in environmental problems like climate change. Therefore, an understandable yet regrettable scenario can arise in environmental problems like climate change even when the agent making the decisions is a unified collective.

4. Conclusion

I argued that the fact that environmental problems tend to be associated with individually negligible effects does not imply that a unified collective has intransitive environmental preferences. I suggested that the PST only appears to reveal the existence of sensible intransitive preferences when the level of certainty possessed by the self-torturer is not taken into account. I argued that an uncertain situation coupled with individually negligible effects can understandably lead the self-torturer or a unified collective into a suboptimal state despite having transitive preferences.

This conclusion is less troubling than Andreou's because it does not undermine existing solutions to some environmental problems that require the assumption of transitive preferences. For example, my argument indicates that a model used to estimate an optimal carbon tax is accurate to the extent that its assumption of characteristically transitive environmental preferences is accurate. However, my argument indicates that aiming for 'optimality' in the sense of balancing the costs and benefits of an activity may not be the best strategy in situations like climate change that are characterized by uncertainty and individually negligible effects. This is because the optimal stopping point is particularly difficult to locate and overshooting this point is particularly costly. This is to say that expected value theory, which is often considered to be the best way to deal with the uncertainties

inherent in climate change decisions (Broome 2012, 117), may in fact not be the best strategy for problems like climate change because aiming to stop polluting at an expected optimal point can easily lead to a regrettable situation. Instead of aiming to balance costs and benefits, we may need to forgo some of the rewards of polluting to avoid overshooting our optimal stopping point in which we would find ourselves in an irreversible and regrettable situation.

Works Cited

Andreou, Chrisoula. 2006. "Environmental Damage and the Puzzle of the Self-Torturer." *Philosophy and Public Affairs* 34,1: 95-108.

Arntzenius, Frank and David McCarthy. 1997. "Self Torture and Group Beneficence." *Erkenntnis* 47,1: 129-144.

Broome, John. 2012. *Climate Matters: Ethics in a Warming World*. New York and London: W. W. Norton & Company.

Brown, Mark B. 2014. "Climate Science, Populism, and the Democracy of Rejection." In *Culture, Politics and Climate Change, How Information Shapes our Common Future*, edited by Deserai A. Crow and Maxwell T. Boykoff, 129-145. London and New York: Routledge.

Kagan, Shelly. 2011. "Do I Make a Difference?" *Philosophy and Public Affairs* 39,2: 105-141.

Quinn, Warren. 1990. "The Puzzle of the Self-Torturer." *Philosophical Studies* 59,1: 79-90.

Tenenbaum, Sergio and Diana Raffman. 2012. "Vague Projects and the Puzzle of the Self-Torturer." *Ethics* 123: 86-112.