Self-trust within trust-betrayal: the whistleblower

Lloyd J. Carr
Philosophy Department
Rivier University
Nashua, NH 03060
lcarr@rivier.edu
Website: http://www.rivier.edu/faculty/lcarr

Abstract:

This paper explores the role of self-trust in a situation in which an autonomous agent is a whistleblower and by blowing the whistle is betraying the trust of others. The paper is organized into three main sections. In the first section I construct a realistic whistleblowing/trust-betrayal scenario in which the agent intentionally both blows the whistle and betrays trust, but in doing so confronts challenges and conflicts at key points. Using four examples of an agent in a position of trust: in a family business, in a military combat unit, in an internship position, and in an academic program, I give the agent reasons to become a whistleblower; I also give the agent equal conflicting reasons not to betray the trust she has been given – trust that blowing the whistle will betray. In the second section I suggest two necessary conditions for trust-betrayal: (i) that, in a full trust relationship in which a trustor trusts a trustee to do an action, and in which the trustor believes the trustee to be relevantly trustworthy to do that action, the betrayer occupies the position of trustee and (ii) that the trustor in this full trust-relationship does not believe the trustee will betray that trust.

In the third section, I argue that self-trust operates at both the intention-forming and the actional stages to keep the agent on-track, so that an overall consistent unit of agency results. Self-trust is shown to operate "positively" with respect to the agent's capability to be responsive to reasons the agent has to blow the whistle, and shown to operate "negatively" with respect to the agent's capability to be unresponsive to reasons the agent has to remain trustworthy. We see that the agent, in trusting herself in this whistleblowing/trust-betrayal unit of agency, does not permit her reasons to remain trustworthy and not betray trust to win out over her reasons to blow the whistle. I next contrast self-trust with a standard conception of self-control and argue that the former, but not the latter, is needed if this unit of agency is to be completed. Finally, I argue that *intra*personal trust can operate on behalf of *inter*personal trust-betrayal (e.g. in the context of whistleblowing) in a morally neutral way, such that independent moral analysis is needed to evaluate the moral status of the initial trust, the self-trust, and the trust-betrayal.

Keywords: Trust, Self-trust, Trust-betrayal, Intention, Whistleblowing, Practical agency, Autonomy

1. Setting up the problem:

It is widely accepted within the literature on trust that trusting other human agents leaves the trustor open to the possibility of betrayal by the trusted.[1] Vulnerability to betrayal is a more troubling aspect of trusting others than the risk that one's trust is disappointed or let down, for betrayal is an intentional violation of trust that takes advantage of trust for its success. Trust-betrayal implies a degree of autonomy and perhaps planning on the part of the trusted; it typically involves a recognition of the potential damage it will do to the trustor and to the future of the trust relationship (indeed, for high-stakes trust relationships betrayal could make future trust impossible), whereas other ways that a trust relationship might be broken need not involve such intention, recognition of damage, or autonomy on the part of the trusted. In letting someone's trust down, the trusted might well be excused – perhaps herself a victim of circumstances or otherwise passive in being relatively not trustable. For example, where person S trusts person T to ϕ, and T in turn trusts institution I to ψ, T might disappoint S's trust by failing to ϕ because I let T's trust down by failing to ψ. In such a case there is no basis for S to claim that T *betrayed* S's trust, for T would not have been an agent in letting S's trust down but more a victim in so far as fulfilling S's trust depended on I's ψ'ing. By requiring that trust-betrayal be intentional, trust-betrayal is necessarily agential.

Two distinctions will help set up the problem I wish to explore. Trust-betrayal is *epistemic* if it takes place between epistemic agents, for example in the area of beliefs and the exchange of information. So, where S and T are members of a pharmaceutical research team and S trusts T to provide accurate information concerning a drug experiment, T epistemically betrays S's trust by intentionally distorting experimental data in an effort to give S false beliefs about the drug in question (even if that effort fails). Trust-betrayal is *practical* if it takes place between practical agents in the area of decision and action. I take the distinction between epistemic and practical trust, and between epistemic and practical trust-betrayal, to be an application to the case of trust of the general distinction between epistemic and practical agency.[2] I will restrict the type of trust-betrayal I wish to explore to practical agency.

The second distinction that helps bring the problem into focus is that between the form of practical agency involving self-trust and the form that does not.[3] We can think of a typical unit of practical agency as an agent moving through three stages, each next based on the previous stage: where ϕ = an action or a course-of-action, the agent moves from an initial desire to ϕ or forming a "judgment" that she should ϕ, to forming an intention or decision to ϕ, and finally to fulfilling the intention by ϕ'ing.[4] In some units of practical agency these transitions do not present the agent with inner challenge, difficulty, or conflict. The agent has no reason to stop or desire to give up; such units happen, as it were, effortlessly and there is no point at which self-trust (or self-distrust) has occasion to operate. Perhaps ϕ'ing is something the agent routinely does, or is an action the agent has practiced to the point of high skill-acquisition, or is simply something everyone does as a matter of daily living. So, for example, we don't normally see the need for self-trust in ordinary hygiene activity such as washing your hands or brushing your teeth even though these (often) count as full units of practical agency in which an agent must transition from judgment or desire to intention and then to action.

In contrast, there are units of practical agency that contain inner challenge and resistance such that an agent requires, to make the transitions from judgment to intention to action, a relation of trust and trustworthiness between her earlier and her later stages of agency. An agent might desire to ϕ or might judge it best (all things considered) to ϕ and then struggle with herself; the agent might experience volitional resistance and conflict, wavering or indecisiveness when it comes to forming an intention to ϕ. Or an agent might form an intention to ϕ but then experience weakness, conflict, lingering doubts and uncertainty, loss of nerve, the appeal of reasons to hold back, temptations to give up, or otherwise find it difficult to get herself to act on that intention when the time comes.[5] Such units of agency, if the agent is to accomplish them, require reasonable self-trust, i.e. the justified self-attribution that "I can and should trust myself" with respect to ϕ'ing, and that "I am trustworthy" with respect to exercising the capability needed to act on the intention to ϕ. In forming an intention to do what the agent desires or thinks it best to do, an agent in trusting herself justifiably believes that with respect to carrying out her plan she is trustworthy in certain capabilities to overcome hesitation and weaknesses, and to resist temptations not to follow through. And, in order to consider herself rationally bound to act on her intention when the moment to act has arrived, the agent must justifiably believe her intention-forming self to have been trustworthy in accepting the prior desire or judgment as the basis on which to form an intention, and trustworthy in reasonable risk-taking and foresight. Self-trust in such challenging cases of practical agency operates as a layer of practical rationality that helps see the agent through the point of forming an intention to ϕ and then through the point of acting on that intention, given a prior desire to ϕ or a judgment that the best thing to do is to ϕ.

The unit of practical agency requiring self-trust I wish to explore is that of betraying another's trust. Betraying a person's or an institution's trust is not, I will assume, an easy thing for an agent to do. For the case I have in mind, the transitions from (i) a judgment, perhaps accompanied by a desire, that betrayal (all things considered) is what the agent should do, to (ii) forming an intention to betray the trust that has been given to and fulfilled by the agent, and then (iii) acting on that intention such that the action is (also, if not primarily) an instance of trust-betrayal, will be difficult for the agent, containing challenges, temptations to give up and reasons not to follow through. Trust-betrayal, we will see, requires the belief that a particular capability or inner strength is required with regard to which the agent normatively expects herself trustworthy in its exercise. My focus, then, is not on the moral evaluation of trust-betrayal (which could be a good or a bad thing depending on the situation), and it is not on the agent's question, "Should I betray the trust I have been given?" I will assume that the cognitive stage of this unit of agency has been concluded: in epistemically non-biased and thorough deliberations, the agent has worked through the reasons for and against – and has overcome any conceptual challenge there might have been in – betraying trust and has come to an all-things-considered judgment that "The best thing for me to do is betray the trust I have been given." The agent now subjectively "*knows"* what she should do. Instead, my interest is on the volitional and actional stages of the unit of agency that come after deliberations have been completed and settled; I want to focus specifically on the role of self-trust as a rational constraint that keeps an agent from backing down when it comes to forming an intention and then when it comes to acting on that intention, given that the agent has strong motivation in the form of compelling reasons *not* to betray the trust that the agent has, prior to betrayal, accepted and with regard to which proven to be trustworthy.

It might be claimed at this point that no one ever judges trust-betrayal as the all-things-considered *best* thing to do, and surely no one ever *desires* to betray the trust they have been given. And, to the degree that the intention to do something is founded on the agent's prior judgment that doing it is what the agent should do (or founded on a prior desire to do it), this would imply that no one ever *intends* to betray trust. Something else must be going on such that trust-betrayal is an unavoidable means to an end or an anticipated but unintended consequence. I believe this view is false. I hope to show that in some situations trust-betrayal is so much a part of whatever else an agent might be trying to do that both judging that trust-betrayal ought to be done and then forming the intention to betray trust are not only possible but in a way uppermost in the agent's mind. The example I have in mind and will use as a model to explore the role of self-trust in trust-betrayal is that of the whistleblower. The whistleblower, I will assume, satisfies the following descriptions:

a) an agent who, prior to blowing the whistle, has been a trusted and a trustworthy member of an organization or institution, e.g. a worker, a family member, a manager, a director, a citizen, a soldier… , who has earned the position of trust s/he is in,

b) an agent who feels a loyalty or obligation toward the organization, institution or individuals that have placed their trust in him or her, and who believes that being trustworthy in his or her position is an important ideal,

c) an agent who believes that the organization or individuals whose trust she has accepted and fulfilled is in many respects an agent of good, neither completely bad/evil nor even a "necessary evil,"

d) an agent who believes that it is exactly the position of trust she has within the organization (or with certain of its individuals) that provides the reason and the opportunity to blow the whistle,

e) an agent to whom blowing the whistle is an intended and an intentional betrayal of the trust given her by the organization or individuals .

No all instances of whistleblowing fit this description; I restrict my topic to those that do. The following whistleblowing scenarios serve as examples of the trust-betrayal I have in mind.

1) Jack has worked his way into an important position within a successful business owned and operated by his family, of which Jack and his family are proud. Because he is now privy to some of the complex ways the company works, Jack becomes aware of an on-going history of successful but illegal tax evasion on the part of his family's business. Jack forms the judgment that, all things considered, the best thing to do is report this systematic evasion to the tax authorities, and this means betraying his family's trust. Given the circumstances, why wouldn't Jack give up on blowing the whistle so that he might remain a trusted family-business member?

2) Jill is a member of a military unit in a combat zone. The unit has trained and has experienced combat together and its members have established strong bonds with each other of loyalty, mutual trust, care and protection. One day several members of the unit tell Jill of a recent action into hostile territory in which they broke into civilian homes, injured the occupants and took a number of valuables, justifying

their activity as their right to some "spoils of war." Jill, however, believes that their actions represent a serious violation of military code. Jill judges that the best thing to do is blow the whistle, realizing that this is betraying her unit's trust in her. Given her strong feelings for her unit and her situation of continued combat, why wouldn't Jill back down when it comes forming the intention to do as she judged she should in order to stay a trusted member of her unit?

3) Jack is a college senior majoring in business, one of the department's best students. He has been placed in the sales division of a local food distribution company as part of the collage's internship program, a successful program Jack's college has developed over many years and through which each year several interns gain immediate employment upon graduation. Jack discovers that the company is distributing a variety of outdated food products as satisfying the official expiration date, and part of his internship involves assuring clients that all items set for distribution are "fresh." In conference with the college's business department Chair, Jack is told that all such companies do this to remain profitable, that the expiration dates have a wide margin of error that minimizes the risk to consumers, and that "causing problems" will not only earn him a poor recommendation toward future employment but would likely jeopardize the department's valuable internship relation with this company. Nevertheless, Jack believes that he should report these food safety violations to the proper government authorities, knowing full well that this is a betrayal of the trust the department and the food service company have given him. Given his situation, why wouldn't Jack fail to blow the whistle so that he can continue to be a trustworthy college intern?

4) Jill, a university student in an academically hard program in which good grades are especially important to advance to graduate school, belongs to a study group. The group not only works well together, they have become close friends who have provided needed academic and psychological support for each other during stressful periods within the pressures and rigors of their program. As final exam week approaches, Jill is told in confidence that someone else in her study group has "found" the up-coming final exam in a particularly difficult mathematics course they are all taking and that the study group will get together to plan how best to use it to maximize their grades without causing suspicion. Jill knows all the arguments that she will get if she tries to convince her group not to cheat: that all students cheat, that grades are not a true measure of learning but a "racket," that a top grade in this particular math subject will highly impress graduate schools, that it is stupid not to take advantage of this "golden opportunity," etc. Jill judges it best, all things considered, to go to the Program Director and make her aware of her study group's plans, realizing that this will betray the trust she has from her study group. What keeps Jill from giving up on blowing the whistle, given her situation?

These are examples of blowing the whistle of which (a) – (e) descriptions are true; they are designed to make whistleblowing an intended (as well as intentional) act of trust-betrayal. True, in each there is the risk of personal costs, perhaps danger, to blowing the whistle that, it is natural to assume, are internal reasons the agent has not to do so; they create conflict for the agent between blowing the whistle and not doing so. But this conflict is not my focus; where whistleblowing is also trust-betrayal, I want to explore the tension an agent faces between betraying trust (by blowing the whistle) and not betraying

trust (by not blowing the whistle) in cases where there is no "easy out" for the whistleblower (W). First, however, I should address possible doubts that betraying trust is what such an agent is intending to do.[6]

It might be argued that condition (e) above (an agent to whom blowing the whistle is an intended and intentional betrayal of the trust given her by the organization or some of its individuals) can't be satisfied and that upon analysis something else must be going on. For example, a doubter might argue that the betrayal of trust is separate from the act of whistleblowing such that what the agent is "really doing" – in judgment, in intention and in action – in blowing the whistle is exposing what the agent believes to be a wrong, it is *not* – in judgment, in intention or in action – trust-betrayal, as if the agent in intending and then doing one thing is neither intending nor doing the other. The case of trust-betrayal I wish to explore, represented in the above four scenarios, is exactly the case in which whistleblowing *is* betraying trust; they are necessarily linked as - so to speak - two sides of one coin, both in the agent's mind and in the agent's reality. In the four examples above, I mean to have trust-betrayal loom so large in the agent' mind that specific attention is given it in the agent's plans; the prior judgment that blowing the whistle is the best thing to do and the subsequent intention to do it are, for the agent, so inseparably linked to trust-betrayal that judging one is best, and forming the intention to do one and then acting on that intention, requires – at least from the agent's perspective – judging the other should be done, forming the intention to do it, and then doing it; the one requires the other. The agent I have in mind has not just made up her mind to blow the whistle, she has also specifically considered its link to betraying trust and has made up her mind not only not to let trust betrayal dissuade her from blowing the whistle, but has made up her mind to go ahead and betray trust, as undesirable as that might be to the agent. For even though the agent is judgmentally sure about what should be done, it is exactly this link between whistleblowing and trust-betrayal in forming the intention and in acting that creates W's main problem in the examples I will focus on and makes this unit of practical agency, as we will see, impossible practically to complete without self-trust.

The correct insight in this argument, I believe, is that a prior *desire* to expose a wrong, if we assume W has such a desire, is not necessarily accompanied by a prior *desire* to betray trust; it would be more natural to assume that a desire to blow the whistle is accompanied by a desire *not* to betray trust; W would like to avoid that part of her course-of-action. In setting up the problem of self-trust within trust-betrayal, then, I will give the agent a desire not to betray the trust she has been given, the trust that she knows that she will not only be betraying by blowing the whistle, but must intend to betray (perhaps with a detailed plan) if her intention is not to be incomplete, ill-formed, and too weak to be the basis for action. The agent's *intention* to betray trust is, in a way, the very thing needed to counter the agent's (strong, it is natural to assume) desire not to do so; an intention to blow the whistle without an intention to betray trust could not counter the agent's desire not to betray trust, for on its own the content of a purely whistleblowing intention does not contain an acknowledgement or recognition of any such desire.

Similarly, W cannot be let off the hook of (e) by arguing that trust is inherently good, thus trust-betrayal is inherently bad, and thus it is always done as the only means – a necessary evil – to the good end an agent wants to achieve; in other words, an agent never *aims* to betray trust, it is rather the means to the goal the agent intends to achieve. There are several problems with such an argument. First, trust is not

always good and trust-betrayal is not always bad. Second, in some instances trust-betrayal might have only instrumental status but this is not necessarily so. We can certainly imagine the example of a spy who has earned the enemy's trust, or an undercover police informant who has earned the trust of a criminal organization, and who when the right time comes specifically aims (and perhaps happily) to betray that trust, and is not much moved by the fact that some military or civic goal is achieved by the trust-betrayal. Trust-betrayal can be its own value to an agent. But even granting that trust-betrayal is just the means to an end, it does not follow that it is not judged the best thing to do or that it is not the object of an intention; by intending to achieve an end it is possible for an agent to intend to do the means, and in judging that a goal should be achieved it is possible also for an agent to judge it best (all things considered) to do the means.

Rather than the means to an end, it might be argued that trust-betrayal is an unavoidable consequence of whistleblowing (in the cases under consideration) and that the agent intends to blow the whistle but does not intend to do or bring about all the undesirable consequences that result from blowing the whistle, including trust-betrayal, that the agent anticipates. Again, it might be and probably is the case in some instances of whistleblowing that there are unintended consequences anticipated by the whistleblower, and that these include betraying trust; but such cases are not the ones that I'm exploring. The whistleblower I have in mind is not only in a position of trust, but betraying that trust is a worry of equal importance to the value of blowing the whistle; it is for the agent the main obstacle to blowing the whistle and the primary reason why the agent must trust herself to complete her unit of agency. W, in forming plans to blow the whistle, realizes that she is equally forming plans to betray trust; and in forming the intention to blow the whistle understands that she is equally committing herself to betraying trust. Betraying trust, in such cases, is not an unintended consequence, it requires the same attention, struggle, and practical reasoning that whistleblowing requires if the unit of agency is to be completed by our agent.

Finally, it might be argued by someone who is troubled by (e) or doubts that it can be satisfied that, in blowing the whistle, W is the one whose trust has been betrayed, W does not betray trust, and that this is what W would (or should) believe in forming the intention to blow the whistle: that she is not betraying trust but that her trust has been betrayed and so, in doing wrong, the organization of which she is a member (or some of its wrong-doing individuals) is not trustworthy and thus no longer deserves W's trust; in other words, W believes (or should believe) that the original trust-relationship has been betrayed by the organization's (or some of its individuals') actions about which W intends to blow the whistle.[7] This argument makes specific assumptions about the structure of the trust complex that (i) might but need not be true – in which case I will focus on whistleblowing cases in which they are not true, and (ii) if accepted without restriction would make a conflict between whistleblowing and trust-betrayal (i.e., not betraying trust) impossible – a conflict we know to be possible. What is the argument against (e) when these assumptions are made explicit?

Where S trusts T to ϕ, and S is an organization of which T is a member in a position of trust, the above argument assumes that the trust relation is symmetrical or bi-directional. So, (S trusts T to ϕ) would imply that (T trusts S to ψ) where ψ'ing = "not doing wrong" and, on a 2nd order, covers the trust S gives to T to ϕ, namely that this trust is not wrong, e.g. not deception. Thus, when S trusts T to ϕ, S and T

believe each other to be trustworthy: T is believed trustworthy to ϕ by S, and S is believed trustworthy to ψ by T. If, then, T's position of trust within S lets T discover an act of deliberate wrong-doing on S's part, it follows that S has betrayed the trust T has given S to ψ, and T has now discovered S to be untrustworthy to ψ. In addition to the assumption that trust in this context is always symmetrical, the above argument assumes that where trust is symmetrical, betrayal in one direction voids the entire trust-complex. Thus, where S trusts T to ϕ, and (symmetrically) T trusts S to ψ, it is impossible for T to betray S's trust if S has betrayed T's trust, for there is no longer any trust to betray. It would follow, then, given that W satisfies (a)-(d), that W cannot satisfy (e) and W never has a conflict between blowing the whistle and remaining trustworthy (i.e., not betraying trust).

(i) While trust can be and no doubt is sometimes symmetrical, it is not necessarily so; the relational state-of-affairs ((S trusts T to ϕ) and (T does not trust S in any respect)) is not impossible, and its description "(S trusts T to ϕ) and (T does not trust S in any respect)" is not a contradiction. Thus, asymmetrical trust is possible. For example, an eye surgery patient, in submitting his vision to the actions of surgical robot, could trust this smart-machine to perform the delicate procedure on his eyes safely; the surgical robot, however, would not be capable of trust at all, much less trust directed to the patient. So, given that both symmetrical and asymmetrical trust-complex are possible, and given that it is possible for whistleblowing to fall into either category, I divide the cases of whistleblowing into (a) those in which there is symmetrical trust between the organization and W, and accept the argument that in blowing the whistle W believes (or should believe) she is not betraying trust but rather has had her trust betrayed, and (b) those in which there is asymmetrical trust from the organization to W such that W believes that blowing the whistle is betraying the trust W has been given by her organization. My interest, then, is only in the asymmetrical case where whistleblowing is trust-betrayal and internally taken to be such by the agent, as represented by the above four scenarios.

(ii) The second problem with this argument that (e) can't be satisfied is that it makes it impossible for such an agent (non-mistakenly) to experience conflict between whistleblowing and remaining trustworthy (i.e., not betraying trust). But, not only is such a conflict possible, we know it to be real. The four scenarios above are not so far-fetched as to be unrealistic, and actual whistleblowers are commonly thought of as "traitors" by the organizations that consider themselves damaged by whistleblowing, indicating that blowing the whistle can be difficult not only because it can be costly to W in the form of reprisals but also because of the fear W has that actual trust relationships will be lost as a result.[8]

I will take the above four examples to show that whistleblowing can be at the same time an act of trust-betrayal, an act that W must specifically intend to do. The problem I want to explore, then, is how self-trust functions as part of rational practical agency in trust-betrayal, using the case of W when blowing the whistle *is* betraying trust and betraying trust is what the agent desires not to do; it's the main reason the agent has not to blow the whistle. My interest is not the agent's reasons to blow the whistle or in whistleblowing *per se*; my interest is the role of *intra*personal trust with respect to the agent's capability to *resist* the reasons she has to stay trustworthy with respect to *inter*personal trust – the capability to be

rationally *un*responsive to reasons the agent has *not* to betray the trust of others by blowing the whistle. To be clear, the agent I am presenting has no direct reason not to blow the whistle as she would, for example, if she had some doubt that the wrong she desires to expose had actually been done, or she had some uncertainty about its degree of wrongness, or believes that someone else would be better suited to blow the whistle. The reason the agent has not to blow the whistle is indirect: she has direct reason not to betray trust and blowing the whistle means betraying trust. Similarly, this agent has no direct reason to betray trust as she would, for example, if she believed the trust she has been given was evil or the trustor deserved betrayal. The reason this agent has to betray trust is indirect: she has direct reason to blow the whistle and blowing the whistle means betraying trust. The problem I wish to investigate, then, is not how self-trust works in a whistleblowing unit of agency; it is how self-trust works in a trust-betrayal unit of agency where the trust-betrayal happens by blowing the whistle. In the next section I set up a general trust-complex and offer an analysis of trust-betrayal, the case of W, within this framework. In the 3rd section I turn to the function of W's self-trust within such betrayal.

2. Trust and betraying trust:

Trust between persons is a more complex relational state-of-affairs than inter*agential* trust.[9] When the trusted (T) is a person, the trustor (S) always risks betrayal in deciding to trust – no matter how well she believes she "knows" T's character or personal qualities; and T's autonomy always allows for the possibility to betray S's trust – no matter how trustworthy S believes T to be or, for that matter, T believes herself to be.[10] If, however, the trusted is a non-human autonomous agent, perhaps a smart machine such as a fully self-driving car or a surgical robot, or a service animal such as a herding or a sight-seeing dog, its autonomy allows for the possibility of *failing* to fulfill the trust it receives, i.e. the possibility of not being trustable, but does not allow for the possibility of betraying that trust. To see why this is so, consider the general structure of interpersonal (as opposed to interagential) trust as the framework within which trust-betrayal becomes possible. I focus more on T (the trusted) than on S (the trustor) because it is T in being trusted to ϕ, not S in trusting T to ϕ, who has the possibility to be the agent of betrayal.

(1) Interpersonal trust is not, or at least not typically, "global" (in the sense of extending to everything the trusted will do). Typically, S trusts T to ϕ where ϕ'ing is a particular action (type or token) or a class of actions related by a particular theme or by an organizing principle. So, for example, S trusts T to pay back a specific loan of money (token), or not to cheat in their romantic relationship (type), or to take care of S's house while S is away on vacation (a set of thematically related actions). It follows, then, that T is trustworthy or not able to be trusted not globally but relative to a particular action or domain of activity. Thus, trust-betrayal, as a form of being not trustable, will be *relative*: it is relative to the particular trust complex to which a particular action or domain of activity is relative. Where S trusts T to ϕ, the correct form of trust-betrayal, then, is not: *T betrays S* (that is, it is not *a person* who is betrayed in trust-betrayal), or even *T betrays S's trust* (that is, it is not a person's trust that is betrayed in trust-betrayal). The complete form of trust-betrayal is: *T betrays (S trusts T to ϕ)*; that is, it is a trust relation, a whole trust-complex, which is violated by trust-betrayal. We see that trust-betrayal is itself a multi-

dimensional relational state-of-affairs: it is relative to S (a trustor), it is relative to ϕ (the action T is trusted to do), and it is relative to ψ (the action by which T betrays that trust). So, where S trusts T to ϕ, T betrays the trust S has given T to ϕ by (i) intentionally not ϕ'ing and (ii) by completing an alternative unit of agency: ψ'ing (where ψ'ing, here the act of whistleblowing, implies not ϕ'ing).[11]

(2) Interpersonal trust, when it is rational, is typically not unconditional. S decides to trust T to ϕ on the basis of believing T to be reasonably trustworthy to ϕ. This belief introduces an epistemic element, and with it epistemic norms, into the practical rationality of interpersonal trust. If such a trust-supporting belief is to be the basis of a trusting that is to count as "good judgment" – trust as a reasonable decision and not a case of "blind trust" or a leap of faith – S's belief that T is trustworthy when it comes to ϕ'ing should be justified. The justifying evidence S has that T can be trusted (i.e. that T is sufficiently trustworthy) to ϕ will always be incomplete, primarily because both knowing another person's character and predicting another person's behavior based on past behavior are always to a degree uncertain. Thus, S's trust-supporting belief that T is ϕ-trustworthy is probable, and in functioning as the basis for trust works to constrain S's trust to a reasonable degree, presumably a degree matching that probability. S, then, will ideally neither be too trusting nor too withholding in trusting T to ϕ. Still, S can never be certain that T will actually ϕ, and the content of S's trust-supporting belief will include the recognition that T might turn out to be, for various reasons, untrustable when the time to ϕ comes about. How, then, should S's trust-supporting belief be described? I believe it is not correct to describe S's trust-supporting belief this way: 'in trusting T to ϕ, S both believes that T is trustworthy and believes that T is untrustable, the former a stronger belief than the latter (or the former more subjectively probable for S than the latter).' And it is certainly wrong to attribute to S, in trusting T to ϕ, the contradiction: 'S both believes and does not believe that T is trustworthy with respect to ϕ'ing.' It seems more true to the phenomenology of trust to describe S's trust, in trusting T to ϕ, as founded on one complex doxic content: [*T can be (to a reasonable degree) trusted to ϕ, but might be unable to ϕ when the time comes*]. In other words, in trusting T to ϕ, S believes that T is sufficiently trustworthy *actually* and *possibly* untrustable only in a blameless and faultless way. It is important to point out that, if this analysis is correct, in trusting T to ϕ, S's trust-supporting belief includes the recognition that T might for various reasons be "rendered" untrustable, but does not include recognition of the possibility that T might *betray* S's trust. Indeed, if S had such a thought, even if it were unjustified and unfair to T, it would surely diminish – if not extinguish – S's trust, and would be a belief that supports distrust.

I am arguing, then, that when S trusts T to ϕ it is not the reasonable acknowledgment on S's part that T could fail to satisfy S's trust when it comes to ϕ'ing that provides part of the context for trust-betrayal, it is the fact that (i) S does not overtly suspect or anticipate betrayal in trusting T to ϕ and (ii) the kinds of untrustability that S's rational trust does recognize are incompatible with – they exclude – the potential of trust-betrayal. The space for trust-betrayal within the trust complex opens up, I am arguing, precisely because S does *not* believe, in trusting T to ϕ, that it exists.[12] Thus, with respect to T's ϕ'ing it is possible for S both to trust and believe that T might not ϕ; it is, however, not possible for S both to trust and believe that that trust might be betrayed.

(3) The evidence that S uses to justify S's belief that T is appropriately trustworthy could have several sources that are independent of T's character; for example, the testimony (perhaps credible or baseless)

of others, or perhaps T "looks trustworthy," or perhaps T is thought "trustworthy by association," or perhaps, as ordinarily happens, S sees that others seem to (continue to) trust T.[13] However, I will assume for the case of W that the bulk of the evidence comes from T herself; for example, a reputation of past trustworthiness in fulfilling a variety of assignments and duties, the manifestation of virtues linked to trustworthiness such as honesty, commitment, perseverance, seriousness of purpose, and the like, a loyalty to the mission of T's organization, and T's verbal assurance to S that T can be trusted to ϕ. Thus, T has both given the impression to S that T is ϕ-trustworthy, and has worked to earn the trust T receives; that is, I will assume that T herself has honestly provided the strongest evidence S might have to believe T is ϕ-trustworthy. And, it is her position of trust that provides T with the reason to blow the whistle in the form of becoming aware of what T believes to be a wrong being done either by or within the organization or institution that trusts T. The situation I am describing, then, is not one in which T is out to destroy S or to seek revenge on S, as if T believes, in being trusted to ϕ, that ϕ'ing is wrong and S is doing wrong in trusting T to ϕ; quite the opposite, T believes that S, in trusting T to ϕ, justifiably expects T to be ϕ-trustworthy based on evidence that T herself has provided that she is; it is not false evidence that T has manufactured for the purpose of betrayal – a ruse on T's part to dupe S. I mean this to indicate that in being ϕ-trustworthy, T has strong reasons to continue to be ϕ-trustworthy; it is not unreasonable, in constructing a model of trust-betrayal, to give T the belief that the ideal of trustworthiness should be an important guide to her actions. Along with the other reasons that bind T to her position of trust, the norm of trustworthiness that T applies to herself makes trust-betrayal in the form of whistleblowing a difficult course of action for T to follow through on. Whistleblowing, then, (i) betrays S's trust in T to ϕ, it (ii) goes against T's ideal of being trustworthy, and it (iii) transforms the good evidence T has provided that she trustworthy to ϕ into the condition on which trust-betrayal (by whistleblowing) can be successful by making it inconceivable to S.

(4) Interpersonal trust implies that T is an autonomous agent, at lease within the range of action T is being trusted to do. When S trusts T to ϕ, S does not directly cause or control T's ϕ'ing such that S makes it impossible practically for T not to ϕ, or else there would be no reason for S to *trust* T to ϕ.[14] T's autonomy has two aspects that are important for my topic. First, autonomy in being trustworthy is 2-sided: on the one hand, it means that T has the freedom to fulfill S's trust as T thinks best, perhaps keeping to the "letter" of that trust or satisfying it in "spirit." It allows T a range of creativity in ϕ'ing, perhaps ϕ'ing with added benefits to S or ϕ'ing in such a way that troubles S even while fulfilling S's trust. On the other hand, when S trusts T to ϕ, T's autonomy makes it practically possible, when the time to ϕ arrives, for T not to ϕ and to leave S's trust unfulfilled. In addition, T's autonomy (along with the opportunity) makes it possible for T to ψ instead of ϕ'ing. Where ψ'ing is whistleblowing, and ϕ'ing is action that represents T remaining a trusted member of an organization or institution, ψ'ing is a betrayal of that trust. While there may be examples of whistleblowing in which T is coerced to blow the whistle perhaps to the point that T's autonomy is in doubt, these are not the cases I wish to explore. I will take it that T both forms the intention to, and then betrays trust by whistleblowing, as a fully autonomous agent.

Second, T's autonomy, if it is to be rationally exercised, means that T is a practical agent who, even though having formed an all-things-considered judgment that she should blow the whistle, is still

confronted with options both (i) at the volitional point of forming an intention to blow the whistle and (ii) at the actional point of carrying out that intention when the time comes. (In what follows, in the formula: *S trusts T to* ϕ, let "S" mean T's organization as the trustor, let "T" mean the potential whistleblower as the trusted, let "ϕ" mean the action of continuing to be trustworthy in carrying out the duties and responsibilities of T's position of trust within S, and let "ψ" mean the act of betraying that trust by blowing the whistle.) Even though T's deliberations leading to the judgment that T should ψ rather than ϕ were presumably difficult to work through, once that judgment is formed it does not "close" the unit of agency to alternatives such that it becomes from here on impossible for T to ϕ. In general, judging what is best to do does not render an agent, for the rest of the unit of agency, unresponsive to reasons to do an alternative action, immune to their continued appeal, or without desire not to do what is judged best to do. So, T (in the model I am describing) experiences (i) volitional conflict and challenge in forming the intention to ψ rather than remaining with the original intention to ϕ for which T has appealing reasons. Likewise, T experiences (ii) actional conflict and challenge when the time comes to ψ rather than ϕ'ing for which T still has tempting reasons. The whistleblowing/trust-betrayal unit of practical agency I am considering, then, is not one in which the agent struggles to form a judgment about what is best to do and then, once these difficult deliberations are completed, is no longer autonomous; more difficulties conditioned on T's autonomy and sensitivity to reasons remain at later points the unit of agency. T's autonomy in this case might seem more a "curse" (in the Sartrean sense) than a blessing, for in forming the intention to ψ it is still possible and appealing for T to intend to ϕ instead; and it is natural to assume (what I have stipulated above) that T experiences the temptation in the form of strong reasons not to form the intention to ψ and instead to reaffirm the original intention to ϕ. T's prior judgment that ψ'ing is the best all-things-considered course of action opens up, it does not collapse, the practical space for volitional struggle in forming an intention to ψ.

It is important to see that the agent's problem of intention-formation is not a question of T re-deliberating the options or of on-going deliberations, as if this would help because, say, initially T did a poor job deliberating in settling the matter about what is best to do; even if T deliberated poorly, re-deliberation or on-going deliberation would not change things, it would just re-start the unit of agency or extend the unit's initial cognitive stage and T would be presented with the same volitional struggle when the time comes to form the intention to ψ. The problem (if it is a problem and not piece of good fortune in our design) is that rational deliberation, no matter how settled, does not – perhaps cannot – reduce potential volitional conflict when it comes to forming an intention to do an action, for the will and its ability to commit an agent to a future course-of-action – at least in the case of humans – is not simply an instrument of our rationality.[15]

Similarly, given that T forms the intention to ψ, T's continued autonomy and rationality generate the same sort of conflict when the time to act comes; T can ψ and thereby carry out the prior intention to ψ, but T still experiences the possibility that she can back down and ϕ instead, going against her intention to ψ, and T has reasons to do so. Practical agency does not happen such that an agent's intention to do something makes it impossible practically for the agent not to act as intended, as if an intention takes away subsequent autonomy and openness to the "voice of reason," and when the moment to act comes, makes the agent a sort of "victim" or a "puppet" of their own prior will. For better or for worse,

an intention does not "pre-commit" or bind your future self to an action to the degree that a reasonable (or even unreasonable) alternative loses all its appeal or that a last minute self-rebellion or spurt (perhaps irrational) of contra-intentional spontaneity is no longer practically possible.[16] In the case of T, recall that at the moment of action there are still strong reasons to ϕ that function as temptations not to ψ. Of course, independent of the continued appeal of reasonable alternatives, an agent by forming an intention to do an action cannot *force* or *make* herself do that action when the time comes; there is no prior guarantee you can give yourself and the unity of a unit of agency is not unity by self-coercion.

What exactly are the options that create T's volitional conflict at the point of intention formation and T's actional conflict at the point of acting? The 2 interrelated options that T has as a potential whistleblower are:

(i) to ϕ – that is, to continue to function as a trustworthy member of the organization or institution that has conferred on T a position of trust. T is being trusted not only (a) to fulfill the duties and responsibilities of her position within her organization or institution, but also and perhaps primarily (b) to (continue to) be trustworthy in doing so. T, then, has multiple reasons *not* to betray trust by blowing the whistle. We can conveniently classify these as: (1) those relating to S, to the institution and its mission to which T has been loyal; these are reasons of institutional expectation represented by the (a) trust; (2) those relating to T's self-interest/prudence in (continuing) ϕ'ing (and which for the purposes of my topic I have neglected); (3) those relating to T's belief in the value of trust/trustworthiness; these are normative reasons both self-applied and institutional expectations represented by the (b) trust. T's other option is:

(ii) to ψ – that is, to blow the whistle and in doing so betray the trust given T by S to ϕ. Whistleblowing is trust-betrayal on 2 levels; it betrays the trust T is given (a) to fulfill the duties and responsibilities of her position in S, and (b) it betrays the trust T is given to (continue to) be trustworthy in (a) activities. The case of whistleblowing/trust-betrayal I'm concerned with – the one requiring self-trust to accomplish – is not, for example, done by T for reasons of financial gain; nor is it done to revenge a prior wrong T believes S has done against T. T, we will assume, has just one reason to ψ: T believes that there is a wrong S, or some of its individuals, is (has been, will be) doing that T's position of trust within S has allowed T to discover. This wrong is, in T's mind, sufficiently grave, and T's situation is such, that T has judged it best, all things considered, to ψ instead of ϕ. In ψ'ing, T desires (has reason to) to blow the whistle and desires (has reason) not to betray trust, but T can't do one without doing the other and T realizes intending to do one requires a specific intention (commitment) to do to other.

We see that T's alternatives at the point of forming an intention and then subsequently at the point of acting on that intention are logically related: ψ'ing implies not ϕ'ing, and ϕ'ing implies not ψ'ing. Thus, T's options are mutually exclusive, an agent cannot do both. While these two alternatives are not necessarily jointly exhaustive (T might be able to avoid both ψ'ing and ϕ'ing through some third alternative), given the assumption that T has accomplished the initial phase of the whistleblowing/betrayal unit of agency in coming to a deliberative judgment that T should ψ, I will further assume that in T's case not ψ'ing implies, in the practical sense, ϕ'ing, and conversely that not ϕ'ing means that T ψ's.

The general structure of rational trust-betrayal, based on the above analysis of W, can be summarily described as follows; given that S trusts T to ϕ:

1) T betrays the trust-complex (S trusts T to ϕ) by: (i) not ϕ'ing, and (ii) by ψ'ing; the scope of T's betrayal is limited to the trust-complex in which T is the one trusted.

2) T betrays (S trusts T to ϕ) as an autonomous practical agent in a unit of agency containing three stages: cognitive - an initial judgment that ψ'ing is the best course-of-action; volitional - forming an intention to ψ; and actional - carrying out the intention by ψ'ing;

3) if T betrays (S trusts T to ϕ), then S believes T is trustworthy to ϕ and does not believe that T will betray (S trusts T to ϕ);

4) if T betrays (S trust T to ϕ), then T (i) falsifies S's belief that T is trustworthy (to ϕ), and (ii) dissatisfies S's expectation that T does ϕ;

5) in betraying (S trusts T to ϕ), T is a rational agent: T has reasons to betray (S trusts T to ϕ) that are sufficient to justify trust-betrayal to T.

I note that it is not a condition on trust-betrayal that the betrayed agent knows or even believes that his or her or its trust has been betrayed, but in the case of whistleblowing the trustor would know (at some point) that his or her or its trust had been betrayed, for whistleblowing is (at some point) a public act even if the betrayed agent does not know who blew the whistle. Given this analysis of trust-betrayal, I now consider the function of self-trust in betraying trust by whistleblowing.

3. Self-trust within trust-betrayal

Intrapersonal trust is in some ways similar to interpersonal trust. We trust others based on our belief that they are trustworthy, not completely but in certain respects. Similarly, to trust one's self is to believe one's self trustworthy, not completely but in certain respects. Such a belief is (or includes) an expectation the self-trusting agent has about herself that, when called upon or needed, she will exercise the capability about which she is believes herself trustworthy. How would self-trust, as so conceived, function in trust-betrayal? To offer my understanding, we must see in what respect the same agent becomes the trustor and in what respect the trusted, and with respect to what capability this agent believes that she is trustworthy.

To review: there are two points in the whistleblowing/trust-betrayal unit of agency at which W experiences resistance and the temptation to give up: the point of forming an intention and the point of acting on that intention. At both points W experiences resistance to blowing the whistle because it is, W believes, a betrayal of trust to do so, and W experiences resistance to betraying trust because W values

being trustworthy and values the trust she has been given by her organization or by some of its individuals. If self-trust is to operate to help the agent overcome these points of resistance, then it must span the unit of agency so as to include both points. In examining how self-trust functions at these two points, I will think of W as a temporally extended agent, a composite of three time-indexed "selves": (i) the cognitive self who, at t_1, deliberates and forms an all-things-considered judgment, "The best thing for me to do is blow the whistle and that means I must betray the trust I have been given"; (ii) the volitional self who, at t_2, forms the intention to blow the whistle and thereby the intention to betray S's trust; (iii) the actional self who, at t_3, carries out her t_2 intention by blowing the whistle and betraying trust. On this model of an agent as an aggregate of multiple selves, self-relations, including self-trust, are conceived to take place within the agent in so far as these selves have forward and backward access to each other, and by such access are able to influence and be influenced by each other, and to coordinate with each other – that is, are able to exercise a capability for forward and backward self-government with respect to each other that results in a unit of agency accomplished by a unified agent. Here is a familiar example of such intrapersonal coordination from another area of action: say you are a "late sleeper," you don't like getting up early in the morning. Today (t_1) you realize that for some reason you need to get up extra early tomorrow morning. Tonight your t_2 self, by forming an intention to get up extra early tomorrow morning (t_3), is able to commit your future t_3 morning-self to getting up extra early, provided that your t_3 sleepy morning-self, who is to carry out your last-night's t_2 intention by getting up extra early, agrees with and accepts the authority that your last-night's t_2 intention-forming self has to commit you to such a "difficult" action – that is, accepts at t_3 your right and your reasons at t_2 to "tell you (t_3) what to do." If at t_3 your sleepy early-morning-self, while confronting the temptation to stay in bed, succeeds in rejecting your t_2 self's authority (perhaps in the morning you dismiss your former self with, "When will I learn; last night it was foolish of me to think that I would actually get up extra early this morning knowing how hard it is for me and how much I like to sleep late."), then (i) you go back to sleep, (ii) your t_2 intention is voided, (iii) your original unit of agency breaks down, and (iv) you experience a degree of disunity as a practical agent that perhaps weakens future self-trust about getting up extra early.

It is important to see that an agent's t_2 self does not have the power to *make* her t_3 self act as her t_2 self wills; an agent's t_3 self can't be (directly) *coerced* by her t_2 self to do her bidding. This is why the relation of trust between the agent's earlier and later selves is needed to bridge these points of agency; not "blind trust" but rather trust that is justified so that the unit of practical agency is rational.

In addition to conceiving W as a composite of three time-indexed selves, for purposes of analysis it will be helpful to think of this unit of agency as running along two parallel tracks: (a) the whistleblowing track – W is rationally motivated to blow the whistle; (b) the trust-betrayal track – W is rationally motivated not to betray the trust she has been given and motivated to continue ϕ'ing. Self-trust, we will see, operates differently at the t_2 and t_3 points in each of these tracks: it functions positively, we might say, in the whistleblowing track with respect to the agent's trustworthiness in being *responsive* to the reasons the agent has to blow the whistle; it functions negatively, we might say, in the trust-betrayal track with respect to the agent's trustworthiness in being *unresponsive* to the reasons the agent has not to betray trust, i.e. *unresponsive* to reasons to remain a trustworthy member of her organization and to

give up forming an intention to blow the whistle or to give up acting on any such intention. As above, S trusts T to ϕ and T betrays (S trusts T to ϕ) by ψ'ing; ψ'ing, then, is separated into:

(a) <u>whistleblowing track</u>		and		(b) <u>trust-betrayal track</u>
t_1: T judges: it is best that I blow the whistle	⇔		t_1: T judge: it is best for me to betray (S trusts T to ϕ)	
⇓			⇓	
t_2: T forms an intention: I commit my (future) self to blowing the whistle	⇔		t_2: T forms an intention: I commit my (future) self to betraying (S trusts T to ϕ)	
⇓			⇓	
t_3: T carries out her t_2 intention: blows the whistle.	⇔		t_3: T carries out her t_2 intention: betrays (S trusts T to ϕ).	

The double-arrows are meant to represent the tight connection between whistleblowing and trust-betrayal both internally and externally as T moves through this unit of agency; this is one package with two linked parts or aspects, it is not two units of agency the agent is simultaneously transitioning. Thus, in what follows the right and left t_1 judgments are each part of one compound judgment, and likewise for the right and left "partial intention" and "partial action." I want to argue that within this 2-track unit, self-trust is a special kind of belief, a normative expectation, that the agent has about herself, namely that she is trustworthy with respect to exercising certain capabilities at certain points as she moves through this unit of agency. To see how self-trust would function, I will apply two general principles of practical agency, which I take to be uncontroversial, if not self-evident:[17]

(1) An agent cannot both form an intention to do an action and believe that she will never do that action.

If you form an intention to do an action, then you believe that you will at some time (try to) do that action (even if it turns out that you don't ever do it).[18] This is a future directed, a forward-looking, principle; it runs from the t_2 intention-forming point to the t_3 actional point of agency, linking an intention to a future action. To violate (1) is to form an incoherent intention and the "unit" of agency would fall apart as irrational.

(2) To act on an intention requires the belief that the intention is still applicable.

If you believe that you are acting on, executing, or carrying out an intention, then you believe that your earlier intention to so act is still in effect; that it is still authoritative and has not been cancelled, overridden, or voided. This is a past directed, a backward-looking, principle; it runs from the t_3 point of action to the t_2 point of intention-forming, linking an action to a prior intention. To believe that your action is carrying out your earlier intention that you believe is no longer in effect might not be irrational in the sense of incoherent, but it makes for a unit of agency motivated by something other than reason.[19]

3.1 Principle (1) applied to the whistleblowing track

It follows from (1) that T could not form the (partial) intention, the volitional commitment, at t_2 to blow the whistle at t_3 if T believes at t_2 that she will not fulfill such an intention and not blow the whistle at t_3. So, a t_2 intention to blow the whistle means that T's t_2 intention-forming self *believes* that her t_3 self "has what it takes" to act on that prior intention and blow the whistle. Such a t_2 belief would not be a prediction about some future event; a *prediction* might epistemically link T's t_2 and t_3 selves but could not offer or provide any practical support to the agent's t_2 self from her t_3 self as "acting on her behalf" as that *belief* about her future actional self would provide to her intention-forming self. Such a forward-looking belief is more in the nature of a *normative assurance* the agent has about her future self, an *expectation* she places on her future self, the content of which is that she can trust her t_3 self to act as she intends because her t_3 self is sufficiently trustworthy to exercise the capability, to summon the needed strength, to do what she earlier intended to do. Because T is motivated to blow the whistle by a reason, at t_2 T must trust her t_3 self to act *in response to* this reason. This is to say: T's t_2 intention-forming self expects her t_3 actional self to be trustworthy in the exercise of the capability to *respond to reason*, to allow the appeal of this reason to move her to action and not at the time of action dismiss it, override it, reconsider it, doubt it, have second thoughts, or diminish the power it had in T's t_1 deliberations. This t_2 expectation directed to T's t_3 self is *normative* in the sense that (i) it should satisfy certain norms: of *rationality* and of *autonomy*. With respect to rationality, that the expectation is reasonable and based on sufficient self-knowledge so that, e.g., the agent does not commit her t_3 self to a course-of action beyond her abilities, or does not hold her t_3 self to an unrealistic standard of trustworthiness. With respect to autonomy, that as part of T's right of self-government, T's t_2 self is within her rights (i.e., it is permissible within the scope of the agent's autonomy) to commit her t_3 self to blowing the whistle by forming the t_2 intention to do so. This t_2 expectation under which T places her t_3 self is also *normative* in the sense that (ii) it is about the norm of trustworthiness with respect to what T's t_2 intention commits her t_3 self to do; that is, that T's t_3 actional self is expected to satisfy the norm of trustworthiness by being moved to blow the whistle for the same reasons that her t_2 self formed the intention to do so. By trusting her future self in this way, T's transition of the intention-forming stage of this unit of agency gains rational support; a transition that would be made more difficult without such support.

3.2 Principle (2) applied to the whistle blowing track

It follows from (2) that at the moment of action, T's t_3 self must *believe* that her earlier t_2 intention to blow the whistle is still in effect, that it now (at t_3) applies to T's actions and has its authority with respect to "dictating" or "guiding" or "directing" what T does. Such a t_3 belief about a t_2 intention would not (only) be about a memory the agent has of the content of an earlier intention, it is more in the nature of a backward directed *normative expectation* that T's t_2 intention-forming self appears trustworthy to her t_3 self in committing her (t_3 self) to a certain course-of-action: blowing the whistle. In accepting the earlier intention as authoritative and applying it to her actions, T's t_3 self trusts her t_2 self to have exercised certain capabilities related to forming an intention and in doing so to have satisfied

certain norms of rationality and autonomy: e.g. that the intention is rationally justified in being based on the t_1 judgment that blowing the whistle is the all-things-considered best thing to do, that T's t_2 self had sufficient foresight with respect to when and how the whistleblowing would take place, that there was awareness of consequences, and perhaps that it included a willingness to accepting responsibility should the action turn out other than anticipated. In a word, that T's t_2 self formed an intention to blow the whistle in a way that is trustworthy in the eyes of her future self whom she has "enlisted" and "burdened" to act on that intention. T's t_3 self, in so trusting her t_2 self, has no reason at t_3 to void, override or reconsider the commitment her t_2 intention puts into effect to blow the whistle, because T's t_2 self appears to her t_3 self to have exercised "good judgment" (i.e., volitional responsibility) and to have formed a rationally justified intention. In trusting her t_2 self, T's t_3 actional self accepts and applies her t_2 intention as "in effect" and binding, and thus believes that she *should* carry out that intention, that she *ought* to blow the whistle as earlier intended and has no reason to hesitate.[20] Such backward-directed self-trust, on this analysis, provides rational support to an agent about to act on a prior intention.

If we combine both principles with respect to the whistle-blowing track of ψ'ing, we see that self-trust is a coordinated mutual trust between an agent's earlier and later selves, each placing the other under a justified expectation of trustworthiness at key points in a difficult unit of agency to exercise certain capabilities needed to transition these points. In forming an intention to blow the whistle, you trust your future self not to let you down by ignoring or dismissing your intention and thereby permitting yourself to be released from carrying it out; and when the moment to act arrives you trust your earlier self not to let you down by involving you in an ill-conceived and unjustified intention. Self-trust, on this analysis, falls under a practical agent's general ability of rational self-management; it functions, we see, to help us negotiate the territory between the constraints of our rationally based intentions and the openness of our autonomy when there are difficulties in doing so. If the agent did not trust herself in this way, it is hard to see how the whistleblowing part of her unit of agency could be accomplished given that it is challenged by the parallel track of trust-betrayal.

3.3 Principle (1) applied to the trust-betrayal track

It follows from (1) that T could not form the (partial) intention to betray (S trusts T to φ) while at the same time believing that she will never act on that intention. T's t_2 intention to betray the trust she has been given means, then, that at t_2 T believes that at some future time (t_3) she will go ahead and betray that trust (by blowing the whistle). As in the whistleblowing track, this belief is not a prediction T makes about what she will do in a certain situation. It is about T's *capability* to do a difficult action, and in this respect it is an expectation that at t_3 T "has what it takes" and will exercise "what it takes" to betray the trust-complex in which she is the one trusted. This is to say, T's t_2 self, in volitionally committing her future self to a course-of-action, *trusts* her t_3 self to fulfill that commitment. This belief, then, is about T's t_3 trustworthiness to do what she earlier (at t_1) made up her mind what she should do and (at t_2) forms the intention to do; that is, it is a normative expectation the agent's intention-forming self has about her actional self. What capability must T's t_2 self believe her future self will be trustworthy to

exercise (at t_3) if T is not to have doubts and be conflicted (at t_2) about her t_3 self's ability to go through with betraying (S trusts T to φ) – doubts and conflict that would undermine forming any t_2 intention to betray (S trusts T to φ)? In what way is T's t_3 self to be trusted by her t_2 self that helps T transition the intention-forming stage of the trust-betrayal track?

I will assume that T knows herself well enough to anticipate at t_2 the following: if at t_3 T remains open to "the voice of reason" and allows herself to be moved by reason, then T will not act on her earlier intention to betray (S trusts T to φ), and this means that T does not blow the whistle; she either (i) allows at t_3 her reasons not to betray trust to override her reason to blow the whistle, or (ii) is caught in a "Buridan" trap of inaction by being equally moved to action by two incompatible sets of reasons: those motivating the need to betray trust (if there is to be whistleblowing) and those motivating not betraying trust. Either way, at t_3 T knows what she *should* do by her t_1 judgment but will not be able to act on her t_2 intention and get herself to do it. (And we recall that T has strong reasons not to betray (S trusts T to φ) that are not made powerless by her t_1 all-things-considered judgment to ψ; trust-betrayal is not something T takes lightly and I have constructed T as an agent deeply conflicted at the t_2 and t_3 points of her unit of agency.) T can at t_3 easily *justify to herself* φ'ing (and not ψ'ing); that's the problem her t_2 self faces in forming an intention to betray (S trusts T to φ). So, the capability needed at t_3 so that T acts on her earlier intention to betray (S trusts T to φ), the capability T's t_2 self must trust her t_3 self to exercise, is to be *un*responsive to (certain) reasons; the capability *not* to be moved by the "voice of reason." I don't mean by this that T neglects the reasons she has not to betray trust, that she refuses to think of them or somehow forces them from her mind. Quite the opposite, it is natural to imagine that the reasons not to betray trust will capture T's full attention at t_3 and impress themselves on her actional self, and that her t_2 intention-forming self anticipates exactly this in *trusting* her future self *not* to act on them. T's t_2 self believes her t_3 self is trustworthy, then, precisely in *not* permitting herself to act on these reasons (not to betray trust), not to "listen" to them in the sense of submitting to their pull. This is not a belief that is a form of wishful thinking: that at t_3 T will somehow not let herself think of them, or will just ignore them, or will no longer have such reasons. I am arguing, then, not that T's t_2 self trusts her t_3 self to be non-rational (and perhaps irrational) by acting – i.e. betraying trust – without reason or without attention to reason; I am arguing that rational practical agency includes the capability *not* to do as one set of reasons dictates (here: reasons to be *inter*personally trustworthy with respect to (S trusts T to φ)) so that the agent can give herself permission to be moved to action by another set of reasons (perhaps weaker than the other in motivating power; here: reasons to betray that trust by blowing the whistle). The capability that T's t_2 self must expect her t_3 self to be *intra*personally trustworthy in exercising, if the interpersonal trust-betrayal track of ψ'ing is to be accomplished, is the strength to *resist* the appeal of certain *good* reasons and to go against them; the capacity, in a sense, to "take sides" against one's self.

It is important to see that this capability to be *un*responsive to a set of reasons is not a matter of simply *not* responding to practical reason as, for example, in the case of someone for whom certain reasons have no appeal because the agent does not understand them, or the case of someone who is, say, so closed off by prejudice that certain reasons can't "get through" to him, or the case of an agent who is so distracted by other things going on at t_3, say the worry of a serious illness, that insufficient attention is

given to the reasons the agent has not to betray trust. Nor is this the standard capability of a rational agent to be "imperfectly" rational and to experience lapses and failures of rationality. And it is not that the reasons to blow the whistle (betray trust) simply outweigh the reasons to remain trustworthy (not betray trust), as if all the agent needs do is make a rational choice between two options (whistle blow or remain trustworthy) according to the weight of the reasons. I am arguing, rather, that practical reasoning can be an agent of resistance; the struggle is *against* the reasons not to betray, it is not *for* the reasons to blow the whistle. This capacity to resist the voice of reason is not an application of critical reasoning skills, a capacity to discover fallacies/flaws in one's reasons (the effect of which would be to neutralize the so-called reasons), and it is not the capacity to rationalize. It is the ability to feel – to rationally experience – the full weight of a set of reasons to φ and then hold back from φ'ing such that without this intentional resistance, the agent would φ. The capability I'm describing is not a form of *akrasia*, of not being able to get one's self to act on the reasons one has, as for example in procrastination. It is the rational strength to "say no" to powerful reasons the agent finds operating "full force," namely, to remain other-trustworthy and not blow the whistle.

This capability T's actional self is being trusted by her earlier intention-forming self to exercise might be thought of as the rational analogue to what is commonly called "self-control," namely the ability to resist a desire to do something; for example, not smoking when you desire a cigarette, or not taking a second helping of dessert when you really want to, or to stop reading when you can't seem to put a good book down.[22] However, I believe that the capability in question is better conceived as a form of practical self-management in which the agent (at t_2) trusts herself to (have the strength to) resist (at t_3) the pull of reasons (and the push of desire) not to betray trust, and to go against the recognition the agent has (at t_3) that she would be acting reasonably and commendably, in her own eyes and in the judgment of others, both by the norms of interpersonal trustworthiness and by prudential considerations, by not betraying trust and not blowing the whistle.

I am suggesting that this forward-directed trust the agent has about her actional self, this expectation about her future trustworthiness with respect to betraying (S trusts T to φ), provides the rational support needed to (help) overcome T's t_2 volitional conflict and potential hesitation about forming an intention to betray trust. Such forward self-trust is *normative* (i) in the sense that it meets (or should meet) (a) norms of reasonableness: for example, that it is not expecting too much of the agent's t_3 self, that the unit of agency in which it functions is not incoherent, that it is not based in self-ignorance or self-deception, and in general that it is not an expectation that the agent function in a way that on reflection is irrational. In addition, it is *normative* in that it meets (or should meet) (b) norms of autonomy: for example, that as part of the agent's right of self-government, T's t_2 self has the right to place her t_3 self under the expectation that she not respond to the reasons she has not to betray (S trusts T to φ), and has the authority to commit her t_3 self to carrying out her t_2 intention to betray that trust; that is, that this is an exercise of – not a self-violation of – the agent's autonomy.[23] It is also *normative* (ii) in the sense that it is about the norm of being trustworthy with respect to carrying out one's plans.

3.4 Principle (2) applied to the trust-betrayal track

It follows from (2) that at t_3 T believes that she is still committed to the course-of-action that will complete her unit of agency, specifically that her t_2 intention to betray trust still applies in guiding and governing her t_3 actions. But at t_3 T has reasons to override this intention or to allow her actional self to ignore it in favor of not betraying S's trust (that she still values). What keeps T from permitting these reasons to give her pause at the moment of action and further to let them cancel her earlier intention, even though she is being trusted by her t_2 self not to do so? It can only be, it seems to me, because T trusts herself with respect to the intention she formed to betray that trust; specifically, that T's t_3 actional self accepts as authoritative and binding (i.e. not cancellable on these grounds) the intention she formed at t_2 because at t_3 T believes – that is, normatively expects – her t_2 self to have been trustworthy in committing her to the course of action she is about to do. In trusting her t_2 intention-forming self, T's rational motivation at t_3 not to betray (S trusts T to φ) is *countered*. It would be wrong, I believe, to describe this as a *weakening* of T's reasons at t_3 not to betray trust, as if this backward-directed aspect of self-trust brought with it new evidence against these reasons; it does not. T's t_3 trust of her t_2 self, her backward-directed expectation of her earlier self's trustworthiness in forming the intention to betray (S trusts T to φ), seems better described as *strength-giving* with respect to not submitting at t_3 to the appeal of those reasons.[24] Applying that earlier intention, then, does not come first; the correct layering, I believe, is that at t_3 T *first* trusts her t_2 self, *then* T's reasons not to betray trust are not responded to, and *last* T submits to her earlier intention. Backward-directed self-trust, on this analysis, gives rational support to T's resistance to (certain) reasons at t_3 in the form of the belief that she is right in permitting herself not to be moved to action by them because to do so would be to go against an earlier intention formed by someone whom she trusts, i.e., believes trustworthy. I am arguing, then, that T's actional self, in trusting her earlier intention-forming self, is reducing the t_3 conflict she faces in acting on that earlier intention to betray (S trusts T to φ); the conflict is reduced by exercising the capability to be unresponsive to their attraction, as appealing (I am assuming) as these reasons not to betray trust are at the moment of action, and it is this capability T's earlier self (trustworthy in T's t_3 eyes) trusts her actional self to exercise.

If we combine both principles with respect to the trust-betrayal track of ψ'ing, we see that (as in the whistle blowing track) self-trust is a coordinated mutual trust between an agent's earlier and later selves, each placing the other under a reasonable expectation of trustworthiness at key points of conflict in a difficult unit of agency; the expectation, that is, to exercise certain capabilities needed to transition these points. In forming an intention to betray trust, you are helped by trusting your future self not to let you down by "listening to" and submitting to reasons not to do so, and permitting yourself to void the authority of that intention and be released from carrying it out. And when the moment to act arrives you are helped by trusting your earlier self not to let you down by involving you in a poorly formed and an irresponsible intention. If the agent did not trust herself in this way, it is hard to see how the trust-betrayal part of her unit of agency could be accomplished given that it is challenged by strong reasons to remain interpersonally trustworthy by not blowing the whistle. Self-trust, on this analysis, falls under a practical agent's general ability of rational self-management in a way that acknowledges – i.e., respects

– the norms of both rationality and autonomy; that is, our rationally justified intentions can't bind us to a course of action if we don't give ourselves permission to let them.

4. Conclusion

It remains to piece together self-trust operating "positively" in the whistleblowing track and working "negatively" in the trust-betrayal track to form the full unit of agency: ψ. There are four linked possibilities with respect to what W faces: (i) blow the whistle, or (ii) don't blow the whistle, (iii) betray (S trusts T to φ), or (iv) do not betray (S trusts T to φ). (i) and (iii) are related as two inseparable parts of one coherent unit of agency: ψ; (ii) and (iv) are likewise inseparably related as two parts of one alternative coherent unit of agency: φ. But W has only 2 sets of practical reasons: those relating to and motivating (i), and those relating to and motivating (iv). Thus, W can transition two irrational (dis)units of practical agency: (a) form the intention to ψ at t_2, in keeping with W's t_1 judgment that ψ'ing is the best thing to do, but at t_3 contradicting that intention practically by φ'ing; or (b) form the intention to φ at t_2, going against W's t_1 judgment that ψ'ing is the best course of action, and at t_3 contradicting that intention practically by ψ'ing. Either way, the unit of agency is irrational (inconsistent) even though the agent at each t_2 and t_3 point conforms to reason: in (dis)unit (a) – good reasons to blow the whistle at t_2 and good reasons not to betray trust at t_3; in (dis)unit (b) – good reasons not to betray trust at t_2 and good reasons to blow the whistle at t_3. Self-trust could not operate in either (dis)unit, and it seems to me that an agent transitioning either one would come away with increased self-distrust, even if the agent got lucky and things eventually turned out "for the best." In each of the two consistent units, φ'ing or ψ'ing, one set of reasons must not be responded to so that the agent can permit herself to be "won over" by the other set. Given the initial judgment that ψ'ing is what the agent should do (as distasteful as trust-betrayal is to the agent), self-trust with respect to trust-betrayal clearly strengthens the agent's ability to transition the challenging stages of ψ. Self-trust works, perhaps paradoxically, to reinforce the practical rationality that the agent *not* let herself be moved by reasons to be interpersonally trustworthy both forwardly at the point of forming an intention and backwardly at the point of action. And self-trust works, more normally, to reinforce at these two points of her unit of agency the practical rationality of the initial deliberations that blowing the whistle is what the agent has the best reasons, and thus ought, to do. Importantly, we see that self-trust also works in a way that does not violate a practical agent's autonomy.

As a final point, I wish to note that I have nowhere argued that self-trust is a good thing; for all we know, W could be completely mistaken and "objectively" should not blow the whistle/betray-trust. W could be a member of a WW II anti-Nazi group of underground partisan resistance fighters, who judges it best, intends to act, and then acts to betray the group's trust by blowing the whistle to Nazi authorities concerning the groups actions. Self-trust, we may assume, would work in such a scenario to help W transition his difficult unit of agency, facilitating a morally bad unit of agency. For a non-whistleblowing example, we might imagine the terrorist suicide bomber having hesitations and misgivings at the intention-forming and at the actional points of his plan to kill innocent people, and for whom self-trust makes it possible for the agent to overcome his practical conflict and transition these challenging points

of his unit of agency. Again, in such a scenario self-trust is not a morally good thing. Also, I have nowhere argued that trust-betrayal is a bad thing. To betray trust that is evil, say trust among a gang of murders, would be morally judged, *ceteris paribus*, a good thing. In examining how self-trust functions in a trust-betrayal unit of agency, I believe it is important to recognize that any moral evaluation must be a separate and independent line of inquiry.

Notes:

1. This claim is qualified by (i) "other agents" because self-betrayal in the case of self-trust does not seem possible, and by (ii) "human agents" because non-human agents, whether animal or smart-machine, even though possessing various degrees of autonomy and able to fail to be as trustworthy as expected, are not capable of betrayal. (The category "non-human" might be too broad here; for my purposes I exclude from "non-human" any super-human beings as, for example, posited by the major religions and any extraterrestrial intelligent life forms as, for example, hypothesized by space science.) Trust-betrayal is, thus, an *interpersonal* relation, as opposed to an *intra*personal or an inter*agential* relation. For an overview of the literature on trust with respect to this claim about trust-betrayal, see McLeod 2011.

2. The distinction between epistemic and practical agency, in general and in application to the case of trust, is not always sharp and perhaps more methodological than reality based.

3. See Carr (2013) for the analysis of self-trust within practical agency that I rely on throughout this paper.

4. "Judgment" is notoriously ambiguous both epistemically and in the context of practical agency. Among its many nuances, it could mean legal "finding" or "opinion" as in "the opinion of the court is …" or "the jury finds the accused …" which are typically understood as the rendering of a judgment. It might mean evaluation as in "the umpire judged the pitch a strike" or critique as in "he is too judgmental when it comes to his children's behavior." "Judgment" might even mean, in some usages, "decision" as in, after much disagreement within a family, the head of household says, perhaps with mock officiousness, "It is my judgment that we will be vacationing at the beach not the mountains." which everyone involved understands to mean "I've decided we'll go to the beach not the mountains." The literature on practical agency and rationality has established "judgment" as a term of art. In line with this use and for the purpose of my exploration here, I take "judgment" to be the outcome of cognitive activity: deliberation and the forming of a *normative belief* in the practical sense of a belief an agent forms about a possible future action, namely that the agent should do that action; for example, "We should use our opera tickets before the season ends." where the opening phrase "I believe that…" is understood. Thus, judgments (in this sense) have truth values and require supporting evidence for their rational justification, and are arrived at by a process of considering the pros and cons of doing an action, and perhaps presented as the conclusions of practical syllogisms. They imply, or include in their content, the agent's physical and psychological ability to do as the agent judges she should do, and are "relative" to a situation, i.e., cancellable if a significant change in the agent's environment or situation takes place. Independent of such a judgment, the agent may also desire or desire not (or

neither) to do what she judges she should. I will take such a normative belief, with or without accompanying desire, to be the initial stage of a unit of practical agency, but before the agent attempts to do what she judges she should, another mental event (or state) – an intention: an act of volition with varying degrees of strength or commitment – is required; I accept the position that belief alone, even when justified, (fortunately, I would add) can't motivate action.

5. Hill (1991), though pursuing a different topic in Chapter 9 "*Weakness of will and character*", offers vivid illustrations of a variety of difficulties an agent might face in moving from judgment to intention to action, all of which he argues can be classifies as "weaknesses of will."

6. In addressing possible doubters in the remaining paragraphs of section 1, I have benefited from reading Bratman (1987), especially chapter 10, "*Intention and Expected Side Effects*" and within it section 10.2 "*The Problem of the Package Deal.*" My position that the agent who intends to blow the whistle also intends to betray trust is based on considerations that Bratman does not address in this chapter, but I do not mean to deny his principle of division for rational intending: (Int (A & B) --> Int (A) & Int (B)), given that the agent believes A and B are actions possible for her to do.

7. I thank Carey Heckman who offered this reaction/argument about (e) in a brief private conversation at the 2013 NNEPA conference, Dartmouth College. The elaboration and criticism of the argument are my own.

8. While I have tailored the description of the four imaginary scenarios of whistleblowing/trust-betrayal in line with my topic, (3) is based on an actual case. See the entry "Whistleblower" in *Wikipedia*, especially section 1.3 "Common reactions," available at http://en.wikipedia.org/wiki/Whistleblower (Accessed 9/20/13.)

9. See Carr (2012) for the analysis of the trust-complex that I rely on here as a framework for trust-betrayal. McLeod (2011) presents a similar analysis in which the trustor always risks the possibility of betrayal in interpersonal trust relations.

10. This, of course, is not true where S and T are the same person; trust-betrayal is not possible in the case of self-trust, though – as in interagential trust – self-trust does allows for various kinds of untrustworthiness (see note 1).

11. I leave unexplored the interesting possibility that when S trusts T to ϕ, T might have a way (creatively) to betray S's trust by ϕ'ing; I don't mean to rule out this possibility, but it would not be the typical case of trust-betrayal.

12. It is not too strong, I believe, to say here that the content of S's trust-supporting belief includes as a component 'T will not betray my trust.' However, the *absence* of a belief that the trusted might or could betrayal our trust seems more true to the experience of trusting others than claiming that we actually form the belief that the trusted will *not* or "could not possibly" betray your trust. The most accurate way to state this might depend on the specific trust relationship; a trustor who was once betrayed and decides to trust again might well have the stronger belief.

13. For the purpose of my topic, I am assuming here that *evidence* justifying belief in T's trustworthiness need not be distinguished from outward indications used to *assess* T's trustworthiness. See, for example, chapter 8 in Kahneman (2011) for studies of facial features and related visual "heuristics" that appear to motivate rapid, automatic, emotionally linked assessments of trustworthiness. For other purposes, however, it would be a mistake to take assessing trustworthiness by outward signs to be evidence justifying trustworthiness.

14. Of course, by trusting T to ϕ, S (typically) gives T a reason to ϕ (as well as a reason to be trustworthy in ϕ'ing); T might be more motivated, or might think herself more obligated, to ϕ as a result of S's trust than would be the case without S's trust. While it is not the case with which I'm concerned, in some trust complexes T might not even know who S is; T has a reason to ϕ in believing that someone or other, or "certain people," are trusting T to ϕ.

15. I take this claim to be uncontroversial both by our ordinary experience of being, sometimes, weak-willed and velleitous even though we have no doubt rationally about what to do, and by the long tradition of philosophical theory arguing that reason and volition are (relatively) independent human capacities. I am not judging, however, whether this is a good or a bad thing. Nor am I claiming that the beliefs about what ought to be done that result from deliberations are not part of the agent's total set of motivations that carry through as the unit of agency unfolds.

16. As in note 15, I take this claim to be uncontroversial in light of the work that has been done by Jon Elster, Thomas Schelling, and George Ainslie (among others) with regard to "Ulysses strategies" of pre-commitment and self-binding, the problem of preference instability, and the problem of sunk costs.

17. "…I take to be…" is meant contrastively; both principles are somewhat controversial and not self-evident in the philosophy of action, intention, and practical reasoning in so far as there are theories that deny a conceptual and a practical connection between intention and belief, in some versions reductively defining "intention" as (part of) intentional action, a "doing." For an overview of such theories, their weaknesses and proposed alternatives see Setiya (2010), especially section 5: *Intention and belief*. Principle (1) is, of course, the point of Kavka's (1983) influential toxin puzzle. Principle (2) is central to Hinchman's (2003, 2009, 2012) explorations of rational agency. The short paragraphs coming after the statement of each principle are my clarifications how each will be applied; they are not intended as attempts to describe their use in the literature.

18. Logically, this claim is neither equivalent to (1) nor does it follow from (1). It is, however compatible with (1) and it seems not unreasonable to assume, as a step beyond (1), that it is true for typical cases of intention-formation, and thus for the case for the unit of agency I am examining. Where (1) is applied to the whistleblowing track in 3.1 below, and applied to the trust-betrayal track in 3.3 below, I make the same move by way of compatibility and assumption.

19. With regard to backward directed self-trust, I have benefited a great deal from several papers by Edward Hinchman (2003, 2009, 2012) made available on his Web Page: http://www.people.uwm.edu/hinchman

20. I am assuming, of course, that T has no reason to begin the unit of agency anew by reconsidering her t_1 judgment; e.g., T's situation at t_3 has not significantly changed from what T envisioned it would be while forming the intention at t_2 to blow the whistle, T has not discovered between t_2 and t_3 that she had made a mistake in her deliberations about the reasons to blow the whistle, and no new turn of events has entered the picture that would cause T to alter her plans.

21. Alfred Mele (1987) describes several strategies and abilities an agent might apply toward "rebalancing" the forces within a set of conflicting motivations, among them a type of "willful inattention" to those motives the agent is attempting to resist in an effort not to act against the agent's better judgment; see especially chapter 4 "Self-Control and the Self-Controlled Person,"pp. 50-61, and chapter 7 "Explaining Intentional Actions: Reasons, Intentions, and Other Psychological Items," pp. 96-108. Mele's primary concern is an account of *akrasia;* because *akrasia* is not the problem W faces either at the intention-forming or the intention-executing stage of the trust-betrayal unit of agency I am investigating, many of Mele's insights concerning ways an agent's might resist unwelcome motives (reasons) do not apply even though W fits Mele's general framework of an agent's capability

for self-control (self-management) when confronted with a tempting incompatible course-of-action to the one judged best and for which the agent has rational motivation that conflict with the agent's motives to act as the agent judged best.

22. This is the common, non-technical, idea of self-control. "Self-control," of course, has broader and more technical meanings in the philosophy of agency (from Socrates to the present) than a person's ability to control (resist) his own desires. See, for example, Mele (1987) on motivational vs. evaluative self-control, pp. 51-55. For the rest of this paragraph see his distinction between skilled and brute resistance, pp. 26-27. Mele's notion of brute resistance, "… it is with a *further intention* that the agent exercising brute resistance forms or retains the intention to do X … The brute resister *intentionally* forms or retains the intention to X." (p. 26), if applied to our case of trust-betrayal, would seem to import the same problem of intention-formation W confronts only now extended to the 2nd order intention.

23. My reasoning here is that an agent's right of self-government would not include the right to violate one's autonomy, for example, by forming the intention to place one's future self into a position of slavery, or a position of confinement or submission that deprived the agent the ability to exercise autonomy. I acknowledge, however, that this is controversial and that there might be situations in which the right of self-government would include the right to violate one's autonomy, even to the extreme of suicide in some situations.

24. Compare Mele's (1987) exploration of self-control, the effect of which seems to shift, redistribute, strengthen or weaken the forces of the agent's motivations, or to alter the agent's situation; it affects the agent's reasons or the agent's world (pp. 58ff). The wrestling analogy (p. 58) is telling: it is wrestling with another. In contrast to this concept of self-control, self-trust – if my analysis is correct – affects the agent's capabilities, for example the agent's strength to resist certain reasons/motives (which have not been weakened), and the assurance with which a person "gives one's self" over to one set of motives rather than another (whose powers have not been altered); by analogy, it is wrestling with one's self.

Bibliography:

Baier, Annette. 1986. "Trust and Antitrust." *Ethics*, 96. 231-260.

Bratman, Michael. 1987. *Intentions, Plans, and Practical Reason*. Massachusetts: Harvard University Press

Carr, Lloyd. 2012. "Trust: an analysis of some aspects." Available at: http://www.rivier.edu/faculty/lcarr

_____. 2013. "Self-trust and self-confidence: some distinctions." Available at: http://www.rivier.edu/faculty/lcarr

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kavka, Gregory. 1983. "The Toxin Puzzle." Analysis, Vol. 43, No. 1, pp. 33-36

Hill, Thomas. 1991. *Autonomy and Self-respect.* Cambridge Univ Press.

Hinchman, Edward. 2003. "Trust and Diachronic Agency." *Nous* 37:1, 25-51 (available at: http://www.people.uwm.edu/hinchman

_____ 2009. "Receptivity and the Will." Nous 43:3 (available at: http://www.people.uwm.edu/hinchman)

_____ 2012. "Narrative and Stability of Intention." European Journal of Philosophy (forthcoming) (available at: http://www.people.uwm.edu/hinchman)

McLeod, Carolyn. 2011. "Trust." In E. Zalta, ed. *The Stanford Encyclopedia of Philosophy*. Available at http://plato.stanford.edu/archives/spr2011/entries/trust/

Mele, Alfred R. 1987. *Irrationality: An Essay on Akrasia, Self-deception and Self-control*. Oxford University Press.

Setiya, Kieran. 2010. "Intention." In E. Zalta, ed. *The Stanford Encyclopedia of Philosophy*. Available at http://plato.stanford.edu/archives/spr2011/entries/intention/

(Fall, 2013)