Forthcoming: Oxford Handbook of Moral Psychology, eds. J. Doris & M. Vargas

Are desires interdependent?

Fiery Cushman and L. A. Paul

Introduction

Some experiences transform us in profound ways. Many people, for instance, are transformed by becoming a parent. How does this occur? Is parenthood like a massive hurricane, which shapes a wide landscape directly and all at once by overwhelming force? Or, could experiences like parenthood transform us the way a spark transforms the landscape—as a small force that only directly touches one dry leaf, but thereby sets off a chain reaction?

To answer this question requires a more precise concept of "transformation" (Paul 2014). It is especially helpful to distinguish two different facets of psychological transformations. Some experiences are "epistemically transformative". These ones impart new knowledge, often of things that we could not have imagined. For instance, a parent might learn what it is like to hold his own newborn child in his arms. As with many epistemically transformative experiences, this one may impart a kind of knowledge that can only be attained through direct experience. Other experiences are "personally transformative"; these give us new preferences (e.g., strong preferences for the welfare of that newborn child, a preference for your child's welfare over your own, etc.). These categories are certainly not exclusive of each other; many transformative experiences are both epistemically and personally transformative. Although the precise natures of epistemic and personal transformations are nuanced, we will adopt a very simple way of talking about them: Epistemic transformations affect our beliefs, while personal transformations affect our desires.

Surely there are some experiences that shape our beliefs, or our desires, like a hurricane. These experiences directly touch many beliefs, or many desires. We assume this is true, and so our concern is not with the hurricane model. Rather, we want to understand whether the spark model is viable. Can a small, local change to one belief eventually cause a cascade of widespread belief revision? And can a small, local change to one desire eventually cause a similar cascade of widespread preference revision? Notice that the metaphor matters: a small spark can only transform a wide landscape if just the right kind of fuel is arranged in just the right way; ten pieces of kindling touching one another can spark a chain reaction, but ten pieces of kindling each separated by an inch cannot. Our question, then, is whether our beliefs and ours desires are typically arranged the right way, like kindling that touches.

With respect to beliefs, the answer is almost certainly "yes". Beliefs are usually interdependent: revising one belief may rationally require revision to other beliefs. This fact is essential to the climax of many mystery novels, when the detective finds a small piece of evidence that forces reevaluation of a wide network of beliefs. Philosophers, like novelists, have long understood the interconnected nature of belief. According to the most extreme

version of this position, belief holism, no representation has a meaning severable from an entire system of representation (Quine 1976, Davidson 1984). The strong claim is controversial, but the more basic insight that beliefs are interconnected is widely accepted.

What about desires? If a transformative experience surgically changes a particular desire, could rationality require that other desires change as well? Are they, in this respect, like beliefs?

The answer is not at all obvious. According to one vision of mental organization, desires are like beliefs in this way, while according to another vision, they are not. By better characterizing each of these visions, we can ask how well each of them aligns with current models of "desire" (what is sometimes called "value-guided decision-making") in the cognitive sciences. This, in turn, will put us in a better position to understand how experiences can lead to profound personal transformations: Always like hurricanes, or sometimes like sparks?

Intrinsic vs. instrumental value

Within psychological research, the concept of "desire" is most naturally associated with our capacity for learning about rewards and making decisions based on that learning. People often organize their behavior in order to maximize subjective reward[1] (Dolan and Dayan 2013). When hungry, people find food rewarding; when lonely, they find companionship rewarding; when tired, they find rest rewarding. Such rewards carry intrinsic value, in the psychological sense. To say that they are "intrinsic" does not mean that we always value them. In some contexts we might not. For instance, when you have eaten enough, food is no longer rewarding. Rather, to say that a reward is intrinsic in this sense means that when it occurs—whenever a person experiences it—that reward is not a means to some further end that is represented psychologically. Rather, the reward functions psychologically as a psychological end in itself[2]. Of course, these rewards may have some further function from the standpoint of natural selection, but they have no further purpose at the level of psychological representation.

Intrinsic reward is to be contrasted with instrumental value. A thing is instrumentally valuable (again, in the psychological sense) because it enables us to attain some kind of intrinsic reward in the long run.[3] For instance, foraging is instrumentally valuable because eventually it allows us to eat, which is intrinsically rewarding. Exercise is instrumentally valuable because it can eventually result in the intrinsic reward of good health. Money is instrumentally valuable because it can eventually be used to attain a variety of intrinsic rewards—good food, good health, and many other goods besides.

The distinction between intrinsic reward and instrumental value has important implications for the nature of personally transformative experience. Recall that personally transformative experiences affect our desires. Now, what if the relevant concept of "desire" here only encompasses intrinsic rewards? In this case, a surgical change to just one of your desires would not have any effect on the other desires you hold. If, for instance, you suddenly find food less rewarding, this does not have any necessary implications for the degree to which you find sleep rewarding, or companionship, and so on. In other words, a spark touching one kind of intrinsic reward has no means of altering others. In this case, if a transformative experience were to affect a wide range of intrinsic rewards, it would have to be in the manner of a hurricane that simultaneously and independently shapes many parts of a landscape.

On the other hand, if the kinds of desires at the heart of personally transformative experiences involve instrumental values, then change to one of them might indeed rationally require change to others.

[1] For simplicity, and following common practice, we will call undesirable experiences (i.e., "punishment") simply as negative reward.

[2] At an ultimate, evolutionary level of analysis, of course, it is likely to be a means to fitness maximization. This is not psychologically represented however.

[3] In this paper, we'll use "intrinsic" and "instrumental" in the psychological sense unless noted otherwise.

Specifically, a transformative experience that directly alters one intrinsic reward will have the effect of altering many "downstream" intrinsic values. If an experience makes you desire food less, for instance, this will change the instrumental value that you assign to foraging, or restaurants, or having a grocery store nearby, and so forth. In sum, altering one desire might indeed rationally require altering others. This introduces the possibility of a spark-like model of personally transformative experience.

Put simply, if the concept of "desire" properly extends only to intrinsic reward, then they are not interdependent. But if desires also include instrumental values, then there will be many rational dependencies among them—change to one desire could compel changes to others.

But are instrumental values a kind of "desire", in the sense relevant to transformative experience? The answer to this question depends both on current psychological models of instrumental valuation, and also on a conceptual analysis of the notion of "desire" relevant to personally transformative experiences.

Two kinds of instrumental value

Broadly speaking, there are two ways in which instrumental value might be represented and used in human decision-making. The first possibility is that it is constructed on-the-fly, in the very moment of formulating a specific plan. In other words, at each moment, a person could derive the long-run expected value of various actions they might perform. These expected values would be instrumental, calculated by multiplying the magnitude of any potential future intrinsic rewards by their probabilities, conditional upon current beliefs about the environment and one's own future actions.

Many of the models of planning commonly entertained by philosophers and psychologists take this basic form. Suppose, for instance, that a person is planning their trip from work back to home. She could call to mind several different routes home, and for each one of these they could consider how well it satisfies their overall goals. Which is fastest? Cheapest? Most scenic? Which allows her to pick up groceries on the way? Eventually she decides to take Main St. home, and this reflects an instrumental value that she has constructed just in that very moment, through the process of planning.

In such cases, it would be an odd choice of words to say that a person "desired" to take Main St. home. This proposition may not be strictly false, but at least it is poorly phrased. It is more natural to say that she "intended" or "planned" or simply "chose" to take Main St. home. In this case, it suggests that the instrumental value of taking Main St. is a fleeting property of a specific episode of planning, constructed on the fly and dispensed with just as quickly. Ordinarily the things we call "desires" are not so fleeting. Presumably this is especially true for the kinds of desires relevant to personally transformative experiences—they are enduring, not fleeting. If an experience merely changes your choices or plans on some afternoon, it is hardly a "personally transformative" experience.

There is, however, a very different way that instrumental value might be represented and used in human-decision making—one that is not at all fleeting. Consider the case of money. Psychologists do not categorize money as intrinsically rewarding, but instead as an object of instrumental value. Infants are not born finding money rewarding, and adults in moneyless societies do not find it rewarding. Unlike food, companionship, or sleep, money has likely not been around long enough for natural selection to encode it as a source of intrinsic reward. Rather, we come to represent money as valuable through a process of learning. Specifically, we learn that it has great instrumental value—by spending it appropriately, we can often obtain other things that we find intrinsically rewarding.

Money, then, is a paradigm example of an instrumentally valued good. Yet, its value is obviously not re-constructed anew, on the fly, each time we choose to spend it. When we spot a spare $5 bill on the sidewalk we do not ask ourselves, "Now, would it be worthwhile to pick this up? What thing of intrinsic reward could I obtain with it? Food? Sleep?", and so on. Rather, we

instantly recognize it as a thing of value, presumably because it has been valuable so many times in the past. This enduring representation spares us the effort of reconstructing its instrumental value from first principles dozens or even hundreds of times per day.

And, of course, it is an entirely natural choice of words to say that a person "desires" to have more money. This stands in contrast to the case of taking Main St. home; it feels less natural to say that the person "desires" to take Main St. These examples hold a more general lesson. When we engage in a process of deliberative planning to derive the instrumental value of some option just for the selection of our next action, or when formulating a specific plan, it seems odd to say that we "desire" the (merely) instrumentally valuable thing. Instead, it seems more natural to say that we intend it or plan to do it. On the other hand, when we assign enduring value to an instrumentally valuable thing due to its common, widespread utility as a means to achieving intrinsically rewarding ends, it seems natural to refer to these enduring values as genuine "desires", just as the intrinsic rewards themselves are.

This point holds broader implications for how we understand personally transformative experiences. To the extent that we often represent instrumental values in an enduring way, there will be a corresponding large set of desires that are interdependent. This property of interdependence is necessary to make viable any "spark"-like model of personally transformative experience.

How often, then, do we represent instrumental values in this enduring way?

The preponderance of cached value

Two basic lines of research suggest that humans frequently and spontaneously construct enduring representations of instrumental value; that is, the kinds of instrumental value representations that we would comfortably call "desires".

The first line of research explores psychological processes of reinforcement. According to early theories of reinforcement, particular actions or behaviors get "stamped in" when they reliably lead to intrinsic rewards (Thorndike 1898). Later theories extended this idea by proposing that similar learning processes could imbue certain actions, objects or events with the status of "secondary reinforcers". Money is a classic example of a secondary reinforcer—it is not intrinsically rewarding; rather, it acquires an enduring status as a valuable thing because an agent learns that it reliably creates opportunities for intrinsic rewards.

In current learning theory, these are often referred to as "cached value" representations (Daw, Niv et al. 2005). Referring to these as "value" representations (rather than "reward") denotes that they have learned instrumental value, and are not themselves intrinsically rewarding. Referring to them as "cached" denotes that the agent stores a representation of their value rather than constructing it on the fly. The main advantage of caching values is cognitive efficiency. It is often computationally expensive to compute the instrumental value of an action, object or event. By substituting a historical estimate of its value based on past episodes, one can achieve large savings in computational effort at a potentially small cost of reward. For instance, if money has typically been instrumentally valuable in the past, it can make sense to simply assume that it remains instrumentally valuable in some present situation rather that calculating anew all the ways one might attain intrinsically rewarding outcomes by spending it.

Research shows that people quickly construct and deploy cached value representations (e.g., Daw, Gershman et al. 2011). If you give people a new learning task—some sequence of oddball colored shapes that they have to click on in order to earn money—within a few dozen trials, they will already have begun to imbue certain shapes and colors with cached value. In other words, in service of cognitive efficiency, they begin to desire instrumentally valuable things almost as if they were intrinsically rewarding themselves. Crucially, however, if the environment changes and these things cease to be instrumentally valuable (for instance, if

a paper currency collapses), their cached value eventually fades away as well.

It is not surprising that people construct cached values with such alacrity and speed, because a large body of research in machine learning and artificial intelligence suggests that adaptive behavior in complex environments requires it. Early AI approaches to games like chess and go foundered precisely on the problem of estimating instrumental value—i.e., of discovering which present move maximizes the probability of an eventual win in the game. These games make it obvious just how burdensome the cognitive demands of "on-the-fly" value estimation can be; and, of course, the everyday environments that humans make decisions in are far more complicated that chess or go. Current breakthroughs in AI gameplay depend in part on the insight that decision-making must be guided by cached representations of value rather than deep, exhaustive search of lengthy decision trees.

A second literature supports the same conclusion that enduring representations of instrumental value are widespread. It begins with the insight, long appreciated by psychologists, that humans construct plans across multiple levels of abstraction. When entering the kitchen in the morning one assembles a plan over representations of coffee-making, cereal pouring, hand washing, etc. Then, upon initiating the coffee-making episode, one moves to a lower level of abstraction, considering water pouring, bean grinding, mug retrieval, etc. Finally, upon initiating mug retrieval, one plans over lower level motor elements. Hierarchical representations of this kind permit large gains in computational efficiency. They allow people to abstract across diverse circumstances, treating all mornings as similar (with respect to the value of coffee) without worrying about the ways in which they often differ (for instance, in the precise location of the coffee mug on the shelf).

Of course, the very logic of such hierarchical representations demands that value is assigned to subgoals, abstracting over the diverse circumstances in which they support some superordinate (and perhaps intrinsically rewarding) goal (Cushman and Morris 2015). Consider, for instance, the act of "tying a bow". This act is organized around a subgoal of, say, securing a knot. Situated at this level of abstraction, the securing of the knot is represented as valuable. And, this value must be enduring so that the subgoal can be reused across many diverse contexts—tying your shoes, trussing a turkey, wrapping a present, and so on. Indeed, the cognitive efficiency of hierarchical planning depends on the "reusable" nature of those subgoal representations. This cognitive architecture requires, therefore, enduring representations of instrumental value.

In sum, humans are designed to promiscuously assign enduring representations to the instrumental value of many events, actions and objects. Because these representations are instrumental they are at least partially interdependent; each of them must be revised when at least some other instrumental or intrinsic values are revised.

Desires are interdependent

Prior philosophical work has established that many of our beliefs are interdependent. In other words, a surgical change to one belief can rationally require the adjustment of other beliefs. Do desires have the same structure? If the concept of desire extended only to those things that we find intrinsically rewarding, it is not apparent why change to any one source of intrinsic reward would rationally require change to any other source of intrinsic reward.

In contrast to this picture, however, the concept of desire extends to many cases of instrumental value. It is perfectly natural to say that a person desires money, a car, a home in a good school district, more exercise, an extra week of paid vacation or, say, tenure. If a person suddenly desired all of these things less (or more), this might well be a substantial enough revision to core aspects of their preferences that it would constitute a "personally transformative experience".

Yet, none of these desires is a plausible candidate for intrinsic reward. Rather, our desire for each of these things is like a form of "cached value" representation—an

enduring representation of instrumental value designed to enable computationally efficient planning. And, because these desires are instrumental, they also depend on other desires. In other words, many of our desires are interdependent. The value of money must change in lockstep with any change to the rewarding things we buy with it.

This property of interdependence opens the possibility of spark-like personal transformative experiences—those in which a relatively small, surgical adjustment of one aspect of our preferences has a ripple effect, rationally compelling the revision of many other aspects of our preferences.

Fagin's folly: Choosing worse by thinking more

Paul (2014) argues that trouble arises for standard methods of decision-making when an experience is both epistemically and personally transformative. If becoming a parent will give us new values that we cannot imagine in advance, how should we go about evaluating whether we want to be a parent at all? Of course, there are some ways of making decisions that still work in such cases: Relying on expert testimony; estimating utilities with a very high degree of uncertainty; flipping a coin. But often we make decisions between alternative actions by trying to imagine, perhaps quite vividly, what things will be like if we choose each of the alternatives. Transformative experiences complicate this particular method of decision-making for two reasons. First, epistemic transformation involves changes that are hard to imagine. Second, personal transformations involve changes to our preferences, and it may be difficult to evaluate future circumstances against one's future (perhaps alien) preferences instead of one's present preferences.

One variant of this challenging aspect of personally transformative experience is personified by the character Fagin in the musical *Oliver!*, adapted from Dickens' Oliver Twist. The aging maestro pickpocket muses about a possible and radical change of course: he will orient his life firmly by a moral compass. In each verse he endorses this abstract commitment, but then begins "reviewing the situation"—i.e., vividly imagining the resulting personal changes. To adopt the general, abstract value of living a moral life would—at least by early Victorian standards—also entail several more specific values: commitment to matrimony and its inevitable compromises; the dignity of honest work; care for one's home and estate; due respect for magistrates and duchesses. Fagin takes these particular subordinate values to follow directly from the superordinate commitment to morally upstanding way of living. (It is, of course, precisely this kind of cascading effect of personally transformation that we have been concerned with.) Yet, while Fagin feels at least mildly attracted to the prospect of being morally upstanding, he is utterly alienated by the more particular values it implies. The kind of person he'd become is just too different from who he is now.

Thus, each of Fagin's verse leads inexorably to the refrain: "I think I'd better think it out again". After several versus of such thoughts, Fagin concludes: "You'll be seeing no transformation".

In its general form (if not its specific content), Fagin's way of thinking is so natural that we can easily overlook its peculiarity. Let us suppose, as the playwright intends, that Fagin has the capacity to adopt quite new moral values for himself. Had the song to have ended differently, in other words, he really could have decided to adopt Victorian standards of a morally upstanding life. For this "new Fagin", matrimony, honest jobs, estates and respectable company would not have felt alienating. Rather, for the new Fagin, these values would be rationally entailed by his superordinate commitment to a morally upstanding life. Moreover, the old Fagin must recognize this fact, because it is their very entailment that alienates him! Yet the old Fagin cannot help but to reject this personal transformation because of the discord between his present values and the implied new ones. His decision-making is like that of a child who refuses to buy big boots for next winter because they don't fit him today.

The problem of evaluating tomorrow's utilities by today's values is discussed in prior philosophical work (for example, Parfit

1987, Ullman-Margalit 2006, Paul & Healy 2018, Pettigrew 2019). Our analysis of "spark-like" personal transformations, however, gives some additional purchase on Fagin's case and the broader category of decision problems that it exemplifies. The central conceit of Fagin's song is that he discovers his alienation from his future self through the process of deliberation by imagination. The song would have been shorter and duller had Fagin merely stated: "In principle I could be moral—but yelch!" What makes the song funny, and Fagin so relatably human, is the mismatch between the higher-order values that he finds attractive and the entailed lower-order values that he finds alienating. Because the process of deriving lower-order values from higher-order values requires cognitive effort this conflict arises through deliberation, and it only grows as deliberation continues. Yet, the result of the deliberation is, at least by one standard, a worse decision. Had Fagin just taken the plunge thoughtlessly and adopted a moral standard, he would might not have only been better, but indeed happier—and, in any event, his new self would not have felt at all alienated from its new values.

More broadly, insofar as humans make decisions about personally transformative experiences by attempting to imagine their future selves, part of what they are imagining is changes to lower-order values entailed by higher-order transformations—in other words, the fires eventually ignited by an initial spark. In such cases we may be especially apt to reject personally transformative change because the network of changes to our values that it entails feels alienating when contrasted with the network of values held by our present selves. Nevertheless, the new self could be a recognizably better one: More moral, more happy, and so forth. On certain theories of rational choice, then, we would be choosing worse by thinking more.

There's a further application of our spark model that's relevant to the possibility of personal transformation. Our spark model permits a more radical change than those we have envisioned, because who we are may be defined not only over our desires, but over a more abstract construal of how sets of desires cohere and affect our behavior.

Consider, for instance, another common personal transformation in higher order values leading to a change in lower order values in a post-transformation self. A father holds his newborn son for the first time, and forms an abstract, higher order desire to nurture and cherish him as much as he possibly can. This adoption of a new higher order, abstract value entails many changes in lower order values; for example, he may no longer value a lucrative job offer that would require him to move across the country and leave his son behind. This, in turn, may cause the father to reconsider a more abstract construal of personal-identity: the father may no longer think of himself as a "career man", for instance. Even if his career-oriented desires remained untouched, their relative place next to family-oriented desires might diminish in a manner relevant to who he takes himself to be.

This is, in fact, an utterly familiar phenomenon. The prospective father might, in fact, be able to appreciate this possibility, at least in the abstract way that Fagin did, as he imagines his future life. Right now, he is committed to the pursuit of career success above all else. He recognizes, like Fagin, that a personal transformation like fatherhood will change what he cares about, affecting not just his love of family, but the place of his career concerns "at the top". This may drive an especially strong feeling of alienation—perhaps he cannot embrace or empathize with that future, less career-driven self. He doesn't recognize that self, or have those values. In this sense, he is deeply alienated from who he'll become once he holds his son.

In general, our sense is that the decision-making challenges of transformative experience become more profound as changes of self become higher-order, or more general. Unless we embrace a strong form of epistemic conservatism, according to which one's present higher order values must always trump one's future (or past) higher order values, we lack a decision rule for this situation. Here it is not quite right to assert we will choose badly by thinking more. Rather, there seems to be no fact of the matter of how to choose at all.

References

Cushman, F. and A. Morris (2015). "Habitual control of goal selection in humans." Proc Natl Acad Sci U S A 112(45): 13817-13822.

Davidson, D. (1984). "On the very idea of a conceptual scheme." Inquiries into truth and interpretation 183: 189.

Daw, N., Y. Niv and P. Dayan (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral ...." Nature Neuroscience.

Daw, N. D., S. J. Gershman, B. Seymour, P. Dayan and R. J. Dolan (2011). "Model-based influences on humans' choices and striatal prediction errors." Neuron 69(6): 1204-1215.

Dolan, R. J. and P. Dayan (2013). "Goals and habits in the brain." Neuron 80(2): 312-325.

Parfit, Derek (1987). Reasons and Persons. Oxford: Clarendon Press.

Paul, L. A. (2014). Transformative experience, OUP Oxford.

Paul, L. A., & Healy, K. (2018). Transformative treatments. Noûs, 52(2), 320-335.

Pettigrew, Richard (2019). Choosing for Changing Selves. Oxford: Oxford University Press.

Thorndike, E. L. (1898). "Animal intelligence: An experimental study of the associative processes in animals." Psychological Monographs: General and Applied 2(4): i-109.

van Orman Quine, W. (1976). Two dogmas of empiricism. Can Theories be Refuted?, Springer: 41-64.

Ullmann-Margalit, Edna (2006). Big decisions: opting, converting, drifting. Royal Institute of Philosophy Supplements 58: 157-172.