Bounded Agency*

Elijah Millgram

Department of Philosophy
University of Utah
Salt Lake City UT 84112

elijah.millgram@gmail.com

June 23, 2019

Recent work on agency has been largely an attempt to characterize the ideal agent, that is to say, not necessarily an agent that is always successful, and not necessarily an agent that is always morally attractive, but in any case an agent that is always and one-hundred-percent an *agent*. It is granted that real-world agency falls short of the ideal, but is either affirmed or presumed that the defective or incomplete agency we all-too-often encounter is to be understood by way of the ideal. To be sure, there is disagreement as to what the anchoring features of ideal agency are, with candidates such as full-fledged commitment to one's actions, knowing what one is doing, and taking on challenges all in the mix.[1]

Here I want to recommend a different approach, one that takes the bounded-rationality research program as a model for investigating agency. Not only is all real-world agency, as I will explain, *bounded*, and not only should we try to make sense of the varied forms of bounded agency on their own terms, without seeing them as deviations from an ideal; we should not be trying to articulate a conception of ideal agency.[2]

*I'm grateful to Christoph Fehige, Svantje Guinebert, C. Thi Nguyen, Constantine Sandis and Aubrey Spivey for comments on a draft; many thanks to the Hebrew University for a Lady Davis Fellowship, and to the University of Utah for support through a Sterling M. McMurrin Esteemed Faculty Award.

[1]See, e.g., Katsafanas, 2013, Korsgaard, 2009, and Velleman, 2015.

[2]For readers who would like to explore parallels with the debate over (non-)ideal theory

1

In the dialect of English encountered in analytic philosophy departments, "agent" has come to be just another word for "person"; here, we want to use it with a more tightly focused sense, which we need to first introduce.

If I am standing on one side of the parking lot, and I kick a ball towards my car on the other side, it might get there... but even if my aim is good, it might not if, for instance, a gust of wind blows it off-course, or if a group of children get in the way, and perhaps pick it up, or if a driver distracted by his search for an empty spot drives over it. Whereas if *I* am going to my car, I will compensate for the force the wind is exerting, and detour around the children, and make sure to catch the eye of that absent-minded driver: in normal circumstances, I will get there *anyway*. Agents absorb various kinds of noise and buffeting from their environment; they stay on track; they exhibit, I will say, *determination*.

Here I am borrowing ideas from recent work by Jenann Ismael; briefly, on her view, what it is to be a self-governing system is to be constructed so as to see a course of action through in a way that buffers and absorbs physical noise: a self-governing system executes its plans resiliently.[3] Extending that view, the buffeting an agent is able to absorb isn't merely physical. A friend has decided that, in the interest of her sanity going forward, an appropriate

in political philosophy, the Fall 2016 issue of *Social Philosophy and Policy*, or alternatively, Enoch, 2018, are recent entry points.

[3]Ismael, 2016, is interested in determination preempting determinism, and so in addressing the traditional free will debate. The notion of what is physically necessary is useful for understanding an open subsystem of the world just in case physical manipulations and interventions produce physically predictable responses. So physical necessity isn't all that useful for understanding self-governing systems.

Unpacking her argument a bit further, Ismael argues (elsewhere, and I won't explain this part of her view here) that modal concepts apply only to subsystems of larger systems, and never to the world-as-a-whole. It follows that the threat of determinism presupposes the effective predictability of open subsystems of the world. That in turn requires internal structure that keeps their behavior reliably covarying with features of the local environment. In self-governing systems of interest to us, the internal state of the system not only changes rapidly, but is, in technical vocabulary I'll appropriate from John Stuart Mill, *progressive* (1967–1991, vol. VII, §III:xv): one of the inputs into the evolving state of the system is its current location in the state space. In self-governing systems, the ways that output depends on input are mediated by the internal state of the system. So the systems of interest to us undermine simple laws that represent behavior as a function of the environment. Such self-governing systems aren't effectively predictable. So determinism is unthreatening.

dosage of task-free leisure time has to be a priority; but every one of the tasks on her bottomless to-do list is accompanied by convincing reasons, and often enough, urgent reasons. If she were unable to resist the force of those reasons, she would be perpetually distracted, discombobulated, and swamped; agency, in her case, consists in part in protecting time in which to do, perhaps paradoxically, nothing.

Perhaps the most memorable picture of agency as I am proposing to understand it comes from Nietzsche, describing how human beings have been brought to have the ability to make and keep promises:

> If we place ourselves at the end of this tremendous process, where the tree at last brings forth fruit... then we discover that the ripest fruit is the *sovereign individual*, like only to himself... [:] the man who has his own independent, protracted will... this mastery over himself also necessarily gives him mastery over circumstances, over nature, and over all more short-willed and unreliable creatures ... those with the *right* to make promises... give their word as something that can be relied on because they know themselves strong enough to maintain it in the face of accidents, even "in the face of fate"... [4]

So the extreme Nietzschean version of an agent, as I'm proposing to construe such a thing here, carries on not only in the face of gusts of wind and passing cars, but despite earthquakes and other disasters; and not only in the face of the sorts of urgent interruptions that impose themselves on my friend's academic and domestic schedules, but even when presented with an offer that, as Mario Puzo's Godfather phrased it, you can't refuse.

Leaving to one side Nietzsche's apparent identification of determination with the will (if you look at the unabridged passage, with *free* will), we can make two further preliminary points. First, agency staying on track isn't necessarily in the service of an objective, that is, a goal or an end. For a politician, say, to continue to live up to his commitment to serve his constituents and his party over the long term, he normally needs to absorb an ongoing stream of political exigencies, which he will do by changing out his goals, not to mention his principles, on a regular basis.[5] This means that

[4]The passage is drawn from *On the Genealogy of Morals*, §II.2; Kaufmann's English rendering can be found at Nietzsche, 2000, pp. 494f.

[5]But then—I am used to hearing—isn't serving his constituents and so on the goal?

we need a more general way of talking about follow-through. Now, before "agent" became merely another synonym for "someone," it meant someone whom you sent somewhere to do things on your behalf, as in the phrase "secret agent". We can tip our hat to the former usage by saying that what an agent carries out is its *mission*.

Second, what aspect of following through on a mission are we centrally after? Reasons for action are almost always *defeasible*: that is, although your conclusion—in the practical case, your evaluation or decision—really does follow from them, additional information or assessments can be nonetheless be apparently compelling grounds from withdrawing it. My friend is right to conclude that she should set aside one day a week for gardening, trail running, and other leisure activities; but her decision can be defeated by, to start off what will prove to be an indefinitely long list, an injured colleague needing her to cover a class, or a hard drive crash, or the need to arrange accommodation for a disabled student... An agent is not only able to compensate for gusts of wind and the like; it is good at absorbing prima facie defeaters—by which I mean now not just considerations that at first glance *appear* to be defeaters, but which, unless they are handled successfully, would *be* defeaters—and at not allowing its mission to be derailed by them.

2

Now we need to fold the notion of bounded rationality into our discussion, and as I introduce it, I'm going to distinguish old-school and cutting-edge

As I'm understanding goals here, they set finish lines one can reach, and they support means-end reasoning: you figure out steps that would reach the goal which, one after the other, you are then to take. (For an account of ends with roughly this shape, see Vogler, 2002.) The psychological correlate of goals or ends is desires, as philosophers nowadays mostly construe them, that is, as built around a representation characterized by one of two directions of fit: if the world isn't as your representation has it, you change the world to make it match your representation, and when it does match, you stop (e.g., Searle, 1983, pp. 7f).

Now a responsible politician will always be updating his somewhat indefinite conception of what it is to serve his constituents and party, in the course of ongoing interaction with them—in something like the way that a responsible teacher continually clarifies to himself what he is attempting to do, or a responsible philosopher clarifies to himself and modifies his conception of philosophy, as he philosophizes. That is, there is nothing like a crisp, stable representation that the world is being brought to match. And there is no finish line, where the constituents have been completely served, and he can go home.

versions.

Sometimes we understand the in-principle-correct way to figure something out, but it won't actually solve our problem, because the procedure would take too long, or is otherwise too resource intensive. For instance, we are told in decision theory classes to choose the option with the highest expected utility; but finding it would typically mean calculating the utility of every option, and satisficing can be a much faster alternative: that is, setting a threshold for what counts as good enough, and taking the first option that comes in over the threshold.[6] Satisficing is an example of a heuristic, viz., a cost-effective and therefore feasible alternative to that in-principle-correct procedure, which is understood to give incorrect or suboptimal answers sometimes, but good enough answers, most of the time or on the most important occasions. Because we have finite minds, and there are bounds to what we can do before a given deadline, we are *boundedly* rather than *ideally rational*—in the old-school way of construing that concept.

The cutting edge of work in bounded rationality drops the assumption that there always (or even mostly) is an in-principle-correct way to figure things out; it avoids using ideal rationality as a reference point. It's all very well to tell people to maximize their expected utility, but as it turns out, just about nobody satisfies the preconditions for having a utility function; for actual human beings, that method isn't well-defined. Or again, we're supposed to compare the utilities of *all* the options, but if those lists of defeaters were genuinely openended, the option space is, again, not well defined.[7] To be rational is to deploy heuristics that won't always give us the right answer, but take into account the costs, computational and otherwise, of figuring things out, and will give us good-enough answers in a timely manner—even when we don't have an ideal with which the heuristics are contrasted.

Here is the observation we need as we turn from bounded rationality to bounded agency: because a heuristic trades off performance on some tasks for speed and cost improvements on others, heuristics perform well in some environments and not in others; there's no such thing as a heuristic that is

[6]For overviews of the early bounded rationality tradition, with discussion of satisficing, see Bendor, 2003, and Conlisk, 1996.

[7]For early work debunking the psychological realism of the conditions for having a von Neumann-Morgenstern utility function, see Kahneman *et al.*, 1982; Millgram, 2005, ch. 10, gives an argument to the effect that if you have a utility function, something is wrong with you. Wimsatt, 2007, is an example of work on that cutting edge.

the right choice regardless of where it's used. In domains where its repertoire of heuristics works well, a boundedly rational agent figures things out more or less rationally; in other domains, as its performance degrades, it is likely to look arational, irrational—or even just plain boneheaded.

3

An agent stays on mission, absorbing (as a character in a Bond thriller once put it) happenstance, coincidences, and enemy action. The resources needed to support determination are expensive, and so any real-world agent faces tradeoffs. And whether or not an agential resource supports determination is location-specific, in roughly the way that bounded rationality is: what keeps an agent on track in one kind of environment, and for one type of mission, will derail it in another.

Laozi tells us: "A weapon that is too strong will not prove victorious; A tree that is too strong will break."[8] We have already contrasted the unrelenting pursuit of an objective with the flexibility about their ends that keeps politicians in the game, and here are a handful of additional illustrations. In transparent institutional environments, it is the upstanding and incorruptible who are able to carry on with their missions, without being sidetracked by temptation and scandal. However, in an organization where functionaries have to be persuaded to do their jobs, and where they can only be paid off or threatened with bureaucratic decisions (not cash or other extra-institutional incentives), getting things done, and so, carrying through with one's mission, requires that an agent be corruptible in the way that his colleagues are. (In the sort of environments I have in mind, if you don't let others do you institutional favors, and make bureaucratic concessions accordingly, you end up completely ineffectual.) In some environments, wealth supports agency; obstacles to one's mission can be surmounted by spending money. But in other environments, the overhead involved in managing wealth turns out to be the impediment that keeps one from sticking with one's plans. In adventure films, it is robust health, athletic ability, and quick reflexes that get the hero to the finish line; in periods of wartime mass conscription, it is the sickly and disabled whose lives manage to stay on track, and the fit whose best-laid plans get left behind in the trenches.

[8] *Daodejing*, ch. 76., trans. P. J. Ivanhoe, in Ivanhoe and van Norden, 2001, p. 200. I'm grateful to Eric Hutton for the reference.

So in general, an agential resource will underwrite determination in some environments and not in others, and a class of special cases of special interest to philosophers is that of different architectures of agency. Some of these correspond roughly to various theories of practical reasoning, and I'll gesture at just a handful of them.[9]

Perhaps the most widely recognized such architecture makes desires—or goals or ends, taken to be the objects of the desires—into the sole initiators of action. An instrumentalist agent of this kind will lapse into apathy or flailing in circumstances where its desires are moot, or irrelevant, or when they trigger only background processes: the *ancien regime* social climber, whose sole concern is his standing amongst the nobility, is left stranded when a revolution does away with the aristocracy; the peasant mobilized entirely by hunger does not know what to do with himself in a post-scarcity economy.[10]

Next, Michael Bratman has blueprinted an agential architecture built out of plans and policies. This sort of agent does not merely pursue whatever goals seem important to him at the moment; once he has adopted a plan, he sticks with it, unless special circumstances arise that would cause him to reconsider, and you would think that this particular design had determination as its very *raison d'être*. Nonetheless, as Candace Vogler has noticed, planning is a successful approach in a managerialist society, where planning is normal, the plans of other people and institutions provide a relatively stable background for one's own plans, and agents have enough in the way of a resource buffer to proceed with their plans. If a would-be Bratmanian agent is so impoverished that trivial unexpected expenses regularly prevent him from stepping through his agenda, or if he lives in the ongoing chaos of a failed state, in which planning is a futile endeavor, the planning approach to life will fail to display as agency, as we are now construing it.[11]

Finally for the moment, agents that are feedback-driven, rather than goal- or plan-driven, can do better in some of the environments in which instrumentalist or Bratmanian agents break down. Such an agent registers when things are, say, going well and going badly, and is disposed to do, respec-

[9]To be sure, not all do: the fictional robots made famous by Isaac Asimov, which act on the instructions they're given, subject to indefeasible side-constraints, instantiate an agential architecture that falls outside discussions of rationality. For treatment of a related personality structure, see Guinebert, 2018.

[10]For some discussion of the problem, see Millgram, 1997, ch. 5.

[11]Bratman, 2007, Bratman, 2018, Vogler, 2002, pp. 106f.

tively, more of the same, or less; so it hill-climbs in the welfare space. An agent of this kind selects new goals on the basis of its affective feedback, and so can continue to operate when its former goals have slipped into irrelevance.[12] Nonetheless, bounded agents of this variety perform well only in some sorts of environments; for instance, a feedback-driven agent can be trapped into addictions. Suppose you're built to do more of whatever feels good, and you come across something that operates directly on the affective signal (nicotine, heroin, fentanyl, cocaine, or even just alcohol); you can end up as a junkie or alcoholic, someone who cares only about that something. But this particular vulnerability is one to which agents that are merely plan- or goal- driven are immune.[13]

We could survey further architectures of agency, but let's recap. The crucial aspect of agency we are emphasizing here is determination: the propensity to carry on with one's mission, despite being buffeted by variation in one's physical, social and other circumstances, and especially, to absorb the reasons to deviate from it that will normally impinge on any extended course of action. The agential architecture which supports that propensity in one sort of environment will fail to do so in others. And that more than suggests that all agency is bounded.[14]

4

Is it really? Christine Korsgaard suggests thinking of agential architectures as the personal analogs of the constitutions of political states; France is on its fifth republic, and I won't even try to count the other forms of government it

[12]For agency of this kind, see Millgram, 2005, ch. 2.

[13]When different architectures of agency generate actions, those actions are likely to themselves be variously structured, in ways which bear the respective stamps of their producers; for an overview of some of the variations, see Millgram, 2010.

[14]Haven't we just characterized agency as essentially extended over time, and so excluded by stipulation agency that only lasts for a moment? Surely not all agency involves *persevering*.

Although the conceptual home of agency *is* diachronic, it does seem to me to have a synchronic limit case. Determination exhibited in a course of action that only occupies a 'specious present'—a moment that is long enough to be noticed, but not longer than that—is a matter not of continuing on with what you are doing, but rather has to do with managing distractions, and in particular the always available second thoughts about the merits of what one is about to do, in that very moment. The determined agent proceeds with his very brief mission, rather than letting the window of opportunity pass.

has traversed; some such revolutions—by which I now mean transitions from one form of government to another, whether violent upheavals or not—were in retrospect inevitable. And here what goes for states goes for people. But even if we allow that any definite constitution will need to be discarded in suitably changed circumstances, what is to say that in principle we could not have a polymorphic agent, one that managed to be an ideal agent precisely by shifting its agential architecture in response to changes in its environment?

We do not want to dismiss agential polymorphism. When human beings shape themselves to fit the highly differentiated slots produced by our division of labor, they adopt agential architectures suitable for one or another expertise. It is not only that the internalized standards and priorities that control their activities shift; deliberation is reshaped to match the requirements of one or another such social niche. For instance, if someone is in middle management, he's probably oriented towards metrics, and he devises strategies for moving them; if someone is a researcher, he is—or ought to be—in the much more tentative business of exploring a terrain; that is, the two ought to decide what to do very differently. Moreover, people are able to move from one such highly specialized mode of agency to another, as when that researcher becomes a dean. (Accordingly, I have elsewhere described human beings as *serial hyperspecializers*.[15]) So we do see agential polymorphism; our question about bounded agency proves to be whether to make sense of the polymorphism we exhibit by constructing a model of an ideally polymorphic agent.

In Woody Allen's mockumentary, *Zelig*, we are asked to imagine a human chameleon, someone who fits in *anywhere*.[16] Now, and this is one way of taking the joke, what Zelig does is not possible: when he is talking to Frenchmen, he speaks French; in Germany, Zelig seems to be able to speak a language he has never learned, or to fake it well enough to fool native speakers. Given basic facts about our limitations—here, it's not humanly possible to speak languages you haven't learned—we are always boundedly conformist; for similar reasons, we always exhibit bounded agency. Pursuing the analogy with old-school bounded rationality would lead us to an argument that one can't reshape one's constitution to suit the demands of any and all environments whatsoever; the train of thought is straightforward enough not to detain us here.

[15]Millgram, 2015.
[16]Allen, 1983.

Let's develop the analogy to cutting-edge work on bounded rationality. We opt for heuristics, you will remember, without necessarily having a conception of ideal reasoning with which to contrast them. Often, ideal rationality as promoted by decision theorists is not merely infeasible but moot, because the preconditions for the ideal being well-defined so rarely obtain in the real world: if your preferences don't induce a utility function, there's no such thing as maximizing your expected utility, and heuristics are thus not always fallbacks for when ideal decision making is too costly or time consuming.

The more we think about an *endlessly* polymorphic agent, the less we understand what we are thinking about, in something like the way that we don't understand what someone's utility function would be, when his preferences are as all over the place as, say, mine. An agent stays on-mission because he has a stake in it, and an agent's stake in his mission is tied to features of his agential architecture; for instance, a goal-driven agent exhibits determination in pursuing his goals, fielding potential defeaters for the steps he is taking toward them, and so on. A polymorphic agent will change out his agential architecture as appropriate in changed circumstances; but when the features of his former agential architecture are no longer present, how are we to understand his stake in his mission being sustained? (Not because he still has his *goals*.)

Adapting our warmup into an illustration, *why* does Zelig try to fit in everywhere? To *really* fit in, you have to be motivated in different ways in different places: in environments where discrimination makes "passing" a reasonable defensive strategy, conformism is motivated by fear; in others—think of the early kibbutzim—it is part of the enthusiastic pursuit of an ideal. If Zelig moves from that hostile society to the idealistic environment of a newly founded kibbutz, and keeps fitting in, in a way that genuinely suits now one, now the other, how can his stake in his mission be the *same* stake? And if we don't understand how it could be, is it the *same* mission? That we can't give a general account of how one's stake in a mission stays the same, when it has to undergo these sorts of substantive changes, tells us that we don't have a coherent conception of determination for endlessly polymorphic agents.

Just as you would not try to understand human language learning and social adaptation by trying to construct a theory of the inner workings of a Zelig, you would not try to explain the agential polymorphism that human beings in fact exhibit by attempting to construct a model of the ideally poly-

morphic agent. Ideal agency would exhibit determination in any environment whatsoever; sometimes it would have to do so through Zelig-like polymorphism; so we should not be using ideal agency as our reference point, when we are trying to understand the bounded agency we encounter in real life.

5

We've been considering an exotic (im)possibility, and it's time to take a few steps back, rehearse how we got here, and say what we've learned.

We haven't given anything like a definition of agency, but we did firm up our conception of it in one direction in particular: it involves determination. And we pointed out that a given agential architecture (or deliberative constitution) will manage follow-through in some environments, rather than others. Thus if an *ideal* agent were one that (among whatever other things) exhibited stick-to-itiveness in any environment whatsoever, it would morph its own agential architecture, as it found itself passing from one such range of environments to another. And in principle, it would have to be ready, willing and able to do so in any of indefinitely many ways, ways that it (and we) can't anticipate.

We then argued that we don't really understand what it would be to be stay determined, across such arbitrary transformations of agential architecture. (That is, however, compatible with recognizing that it has been managed in one or another case.) Just as steam engines, long ago, had governors, different agential architectures rely on different controlling elements; we mentioned desires, plans, and affective feedback loops, but there must be indefinitely many types of controlling element doing their jobs in the indefinitely many arbitrarily different deliberative constitutions needed to keep an agent functional in the endless and as-yet-unimagined circumstances it might yet encounter. Without knowing what those controlling elements are and how they work, we don't have the wherewithal to make sense of an endlessly polymorphic agent staying on point. If we can't know what determination looks like in the endlessly polymorphic agent, we had better not insist on understanding ideal agency before trying to make philosophical sense of the bounded agents we are and meet on the street.

Evidently, the next stage of a successful investigation of agency will proceed in the piecemeal manner of old-time botany and zoology. Given that all agency is bounded, we can anatomize first this type of bounded agent,

and then that one, in each case looking at how it is that it remains on track, doesn't give up, and completes its mission in the environments for which it is suited. As with bounded rationality, where the question is not what rationality *tout court* is, but what it can and ought to be for us, the question at hand is, What are the various forms of bounded agency that will allow you to, as the Marines recruiting slogan has it, be all that you can be?

References

Allen, W., 1983. *Zelig*. MGM, Santa Monica. Produced by Jack Rollins, Charles Joffe and Robert Greenhut.

Bendor, J., 2003. Herbert A. Simon: Political scientist. *Annual Review of Political Science*, *6*, 433–471.

Bratman, M., 2007. *Structures of Agency*. Oxford University Press, Oxford.

Bratman, M., 2018. *Planning, Time, and Self-Governance*. Oxford University Press, Oxford.

Conlisk, J., 1996. Why bounded rationality? *Journal of Economic Literature*, *34*, 669–700.

Enoch, D., 2018. Against utopianism. *Philosopher's Imprint*, *18(16)*.

Guinebert, S., 2018. *Hörigkeit als Selbstboykott*. Mentis, Paderborn.

Ismael, J. T., 2016. *How Physics Makes Us Free*. Oxford University Press, Oxford.

Ivanhoe, P. J. and van Norden, B., editors, 2001. *Readings in Classical Chinese Philosophy*. Hackett, Indianapolis, 2nd edition.

Kahneman, D., Slovic, P., and Tversky, A., 1982. *Judgment under Uncertainty*. Cambridge University Press, Cambridge.

Katsafanas, P., 2013. *Agency and the Foundations of Ethics*. Oxford University Press, New York.

Korsgaard, C., 2009. *Self-Constitution*. Oxford University Press, New York.

Mill, J. S., 1967–1991. *Collected Works*. University of Toronto Press, Toronto.

Millgram, E., 1997. *Practical Induction*. Harvard University Press, Cambridge, Mass.

Millgram, E., 2005. *Ethics Done Right: Practical Reasoning as a Foundation for Moral Theory*. Cambridge University Press, Cambridge.

Millgram, E., 2010. Pluralism about action. In O'Connor, T. and Sandis, C., editors, *A Companion to the Philosophy of Action*, pages 90–96, Wiley-Blackwell, Oxford.

Millgram, E., 2015. *The Great Endarkenment*. Oxford University Press, Oxford.

Nietzsche, F., 2000. *Basic Writings of Nietzsche*. Random House, New York. Edited and translated by Walter Kaufmann.

Searle, J., 1983. *Intentionality*. Cambridge University Press, Cambridge.

Velleman, J. D., 2015. *The Possibility of Practical Reason*. Maize Books, Ann Arbor, 2nd edition.

Vogler, C., 2002. *Reasonably Vicious*. Harvard University Press, Cambridge, Mass.

Wimsatt, W., 2007. *Re-Engineering Philosophy for Limited Beings*. Harvard University Press, Cambridge.