Journal of
Cognition
and
Neuroethics

Free Will and Autonomous Medical Decision-Making

Matthew A. Butkus
McNeese State University

Biography

Dr. Matthew A. Butkus earned his undergraduate degrees from Georgetown University (German and Philosophy) and the University of Pittsburgh (Psychology). He earned his graduate degrees from Duquesne University (MA in Philosophy and PhD in Health Care Ethics). He is currently an Associate Professor in the Department of Social Sciences at McNeese State University. His prior academic and clinical appointments include Chatham University, Mercy Hospital–North Shore Campus, the CRISMA Research Group in the Department of Critical Care Medicine at the University of Pittsburgh Medical Center (Clinical Research, Investigation, and Systems Modeling of Acute Illness), and St. Francis Health System.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Butkus, Matthew A. 2015. “Free Will and Autonomous Medical Decision-Making.” *Journal of Cognition and Neuroethics* 3 (1): 75–119.

Free Will and Autonomous Medical Decision-Making

Matthew A. Butkus

Abstract

Modern medical ethics makes a series of assumptions about how patients and their care providers make decisions about forgoing treatment. These assumptions are based on a model of thought and cognition that does not reflect actual cognition—it has substituted an ideal moral agent for a practical one. Instead of a purely rational moral agent, current psychology and neuroscience have shown that decision-making reflects a number of different factors that must be considered when conceptualizing autonomy. Multiple classical and contemporary discussions of autonomy and decision-making are considered and synthesized into a model of cognitive autonomy. Four categories of autonomy criteria are proposed to reflect current research in cognitive psychology and common clinical issues.

Keywords

Cognitive psychology, neuroscience, medical ethics, autonomy, heuristics, backstage cognition, decision-making

Since its inception, medical ethics has concerned itself with balancing several key concepts—the patient's best interest, both psychosocial and medical; the patient's legal rights and autonomy; the authenticity of the patient's decision, i.e., narrative concerns that the patient's choice be reflective of her values, etc. As is the case with any pluralistic system, these concepts are complementary at times and conflicting at times. Significant efforts to determine just how to proceed in any given case result, both in academic circles, in which theories clash and value structures rise and fall, as well as in clinical cases, in which academic language gives way to clinical context and lives hang in the balance.

These concepts of autonomy and authenticity have dominated ethical thought for several decades, and have been given significant, if not complete, weight in many theories. Autonomy is seen by many as a deontological norm—an absolute right and duty in some models, a *prima facie* duty in others. Its value and moral weight are understood as being *a priori*—it is not contingently valuable or worthy simply as a means to some other end. The purpose of this paper is to explore this concept of autonomy, and to see how it is modified by knowledge from multiple fields.[1] Philosophy certainly offers

[1] I first explored this critique of autonomy in light of depressive illnesses and the decision to forgo medical treatment in my doctoral dissertation (see Butkus, M. A. (2006). *Depression, Volition, and Death: The*

compelling accounts and definitions, but a fundamental question arises: what does the concept mean in light of what we have learned from fields like cognitive psychology and psychiatry? Philosophy and ethics have debated ‘modifiers of the voluntary’ for a long time, but these concepts of coercion generally are predicated on conscious awareness and experience.

A more complete model of cognition notes that significant thought processes occur at levels which we are only beginning to understand. These influences are non-conscious: they stem from a collection of processes outside of our conscious awareness. How, therefore, can we exercise control over or appreciate the influence of elements of which we aren’t aware? Many models fiat the ability of the moral agent to choose amongst alternatives—these models seem to be less compelling in light of what we know and understand from other disciplines. In fact, the more we learn about the brain, the more homuncular they seem—it is almost as if they argue for a little man sitting in our brains, selectively choosing what will influence us to act. These models are untenable—any conception of autonomy must include an appreciation for cognitive elements outside cognition, which potentially bias us in ways that are inauthentic. In upholding choices that may be skewed or inauthentic, we undercut any meaningful sense of autonomy. A proper sense of autonomy, therefore, is much more deterministic and less ‘rational’ than modern models suggest. As such, greater care is necessary in assessing competence to forgo treatment—quite simply, current popular models allow for more bad decisions with fatal consequences, a reality antithetical with the stated and implied purposes of ethics in medicine. We destroy that which we would protect in a decision which may be the last choice the patient ever makes. If we genuinely care for our patients, we ought to help them reach meaningful choices, instead of fiating an empty and ill-defined autonomy.

Case Study

William R. is a 45 year-old man with end-stage renal disease. He is dialysis-dependent and requires treatment three times per week. In his last hospitalization, he explained to his treatment team that he no longer desired to receive dialysis, maintaining that he felt

Effects of Depressive Disorders on the Autonomous Choice to Forgo Medical Treatment. Pittsburgh, PA: Duquesne University. doi:10.13140/2.1.3236.9284). That research and analysis strongly informs this work. Recent arguments in medical ethics have also explored the impact of mental disorders and their implications for the free will debate (see Meynan, Gerben. 2010. “Free will and mental disorder: Exploring the relationship.” *Theoretical Medicine and Bioethics* 31: 429–443.; Müller, Sabine, and Henrik Walter. 2010. “Reviewing Autonomy: Implications of the Neurosciences and the Free Will Debate for the Principle of Respect for the Patient’s Autonomy.” *Cambridge Quarterly of Healthcare Ethics* 19: 205–217).

it would be too burdensome for him to continue. An ethics consult was called, and the consult team met with the patient for over an hour, discussing his understanding of what forgoing treatment would mean and his reasons for electing not to continue dialysis. He described his personal history, in which friends of his became dialysis dependent and were unable to continue with their hobbies and personal interests. He disclosed that he found the treatments prevented him from eating out, socializing, and enjoying other activities. He understood that absent his treatments, his physical state would deteriorate, culminating in his death in a matter of weeks.

Throughout his interview, the consult team did not find any immediate reason why he could not exercise his personal autonomy and forgo treatment. The consult was written up elucidating the reasoning for supporting the patient's decision and the team's recommendations were followed. William died the following week.

Critiques of Autonomy and Classical Models

William's case is an example of a classic issue in medical ethics—the ability to forgo medical treatment is quite likely *the* most common and accessible example of medical ethics to the general public, with representations in popular television, movies, and other mass media. A patient's ability to act autonomously is rightly praised—individual liberty is highly valued in Western society and if our ability to act is going to be curtailed, we require a significant level of justification for doing so.

This is not to suggest that autonomy is not without its detractors—the question is raised as to whether we have overcorrected from the paternalism of the past, in which physicians would routinely substitute their own preferences for those of their patients. Autonomy has been criticized from feminist philosophy and sociological viewpoints,[2] for

[2] Donchin, Anne. 2001. "Understanding autonomy relationally: Toward a reconfiguration of bioethical principles." *Journal of Medicine and Philosophy* 26 (4): 365–86; Homan, Richard W. 2003. "Autonomy reconfigured: incorporating the role of the unconscious." *Perspectives in Biology and Medicine* 46 (1): 96–108; Jennings, Bruce. 1998. "Autonomy and difference: The travels of liberalism in bioethics." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 258–69. Upper Saddle River: Prentice Hall; Lane, Robert E. 2000. "Moral blame and causal explanation." *Journal of Applied Philosophy* 17 (1): 45–58; Light, Donald W., and Glenn McGee. 1998. "On the embeddedness of bioethics." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 1–15. Upper Saddle River: Prentice Hall; Parks, Jennifer. 1998. "A contextualized approach to patient autonomy within the therapeutic relationship." *Journal of Medical Humanities* 19 (4): 299–311; Roessler, Beate. 2002. "Problems with autonomy." *Hypatia* 17 (4): 143–62; Tauber, Alfred I. 2003. "Sick autonomy." *Perspectives in Biology and Medicine* 46 (4): 484–95; Wolpe, Paul R. 1998. "The triumph of autonomy in American bioethics: a sociological view." In *Bioethics and Society: Constructing the Ethical*

instance, and arguments for softer forms of paternalism have been recently proposed (Conly 2013). Despite these critiques, autonomy remains highly valued in both law and philosophy, and we still maintain a high standard for valuing other principles above it. The issue at hand is whether this standard is artificially high—that is, is it higher than warranted given more naturalistic explorations of the phenomena of consciousness and decision-making. Does it make sense for us to reappraise both the value of autonomy as well as the criteria defining it?

Classically, autonomy has been linked with human reason—for theorists like Aristotle, Kant, and Descartes, reason and rationality are essential defining characteristics of humanity. This classic model of autonomy involves five assumptions about rationality, including literalness, logic-dependence, conscious experience, disembodied transcendence, and essential emotional disconnectedness. Contemporary research has challenged or debunked these assumptions, yielding a model of rationality that is dependent on metaphor, metonymy, inferential reasoning, and unconscious processing, and which is fundamentally connected to and influenced by emotion (Lakoff 1999). Further, this model is known to skew perceptions of new evidence, to have limits in scope, and to be contextualized (Evans & Hollon 1988; Miller & Moretti 1988). Ultimately, this empirical cognitive model defies rationalist claims. This does not make it easy, however, to abandon classic notions of radicalized autonomy—there is still a visceral appeal to the idea that I am in full control of my thought process and the actions that result from it. However, if we want to be honest and move towards a sense of autonomy that matches up with the available data, we must become much more aware of the role of the unconscious and backstage elements of our cognition. Continuing to insist that medical autonomy reflect classical and rationalist models of cognition is dangerous—it promotes an ideological model divorced from actual decision-making. Human thought is much more complex, reflecting deductive, inductive, and abductive reasoning influenced by unconscious and backstage elements *and* streamlined by a number of cognitive heuristics hardwired by evolution. The recognition of these unconscious backstage elements has required a reimagining of the concept of freedom and autonomy (Hájíček 2009; Levy 2003; Shepherd 2012).

Cognition is not a single-stage process—there are many levels of organization in the brain, and they interact with each other in many ways which are open to influence. Conscious thought—the result of myriad physical and social interactions, is also a construct; a concatenation of many different types of cognition, operating in conscious

Enterprise, edited by Raymond DeVries and Janardan Subedi, 38–59. Upper Saddle River: Prentice Hall).

and backstage capacities. Backstage cognition involves a variety of related concepts, e.g., reflexive thought patterns with affective and behavioral components, generation of novel meanings for situations and objects from the mental assembly of other situations and objects, and distinctions between algorithmic and heuristic thought. Our concern is not with the conscious elements of cognition, as conscious phenomena are predicated on deeper phenomena. We cannot have conscious experience without deeper structures, much as we cannot build a castle before constructing its foundation. All of the myriad sense data we take in initiate complex activation pathways, associating current stimuli with previous experiences, affective data, and other valence structures. These deeper cognitive phenomena are not simplistic processes—they are layered, quite complex, exceptionally fast, and quite independent of our volition (Ashcraft 1994). Their automaticity belies their complexity—just as complex physical responses can be initiated without volition, so too we should recognize that our cognitive processes can be induced to action. An environmental trigger can give rise to the activation of many complex systems—a particular memento can trigger complex memory and affective components with corresponding behavioral components (Smith 1997). For instance, I may pass a photograph of my grandfather, which triggers a series of memories (living with my grandparents, visits, holidays, advice given to me, etc.), eliciting specific affective responses (sorrow at his passing and resolution to fulfill promises made to him), and culminating in actions and behavioral changes. None of these responses were necessarily *chosen* by me—they are all direct results of the environmental stimulus; further, this same stimulus can affect me well after I actually encounter it—my *memory* of the stimulus can provoke identical psychological and behavioral responses.

What is more, these backstage processes are also able to introduce errors into cognition—the way we perceive the world is dependent upon a variety of factors, some within our control, some well outside control. A requisite part of accurate cognition is appreciating and understanding when we are making choices based upon the indeterministic elements within our control and the deterministic elements lying outside our volition or awareness.

Automaticity is a significant element of cognition—a variety of processes simply occur without volitional cueing.[3] Bargh understands automatic cognitive processes

[3] The simplest means of demonstrating this is by asking the question "What is the first thing you think of when I say the words 'white bear'?" The normal reaction is to call to mind immediately an image of a polar bear—this was not a voluntary process, however, in that had the words pointed to some other cognitive target, you would be free to think of myriad other things instead of white bears.

to occur “reflexively whenever certain triggering conditions are in place; when those conditions are present, the process runs autonomously, independently of conscious guidance” (Bargh 1997). This can refer both to physical processes—such as navigating an automobile while thinking of something entirely different—as well as cognitive processes—such as references to white bears cueing the imagination of polar bears. Isen and Diamond clarify Bargh’s model, noting that automatic processes are best understood as a ‘parallel process’—they do not take up cognitive processing resources (attention or effort), so they can occur parallel to other cognitive processes which do require these resources (Isen 1989). Because it does not tax cognitive resources, automatic processing can be performed much more rapidly and earlier than other types of processing. This may explain our ‘gut instincts’ in certain situations—our full processing has not yet finished, leaving us with only a general impression of necessary action. Berkowitz notes that the deterministic model suggested by automaticity is frequently undervalued by many people—there is a frequent visceral objection to the idea that our cognitive processes are heavily influenced by environmental determinants (Berkowitz 1997). These can be manifested as objections to experimental results or methodologies or as appeals to the indeterministic claims of folk psychology. Berkowitz suggests that, if nothing else, “Persons interested in gaining a truly adequate understanding of the complexities of human conduct should at least adopt a healthy skepticism toward the assumption that conscious processes are necessarily involved in all human behavior” (Berkowitz 1997, 85). As much as the average moral agent would like to dismiss them, unconscious and preconscious processes can be powerful determinants, and not just modifiers, of the ‘voluntary.’

Preconscious processes develop as the result of conditioning—we develop patterns of psychological responses to stimuli. As is claimed by behaviorist thought, we make associations between stimuli and psychological responses, facilitating future responses along those same psychobehavioral lines. It becomes easier for stimuli to elicit behavioral, emotional, and motivational responses in us, producing automatic cognitive processing. Initially these responses require work, but like other recurring responses, the amount of conscious effort they require consistently decreases to the point where they require no conscious processing at all (Bargh 1997). This has serious ramifications—it means that if we encounter a particular cognitive trigger, we can initiate goals, motivations, and resultant behaviors automatically. Absent volitional control, we may not necessarily be able to control the kinds of thoughts and actions that result. In a clinical setting, for instance, a particular diagnosis may be an emotional trigger for a variety of subsequent thought processes and associations. The mere word ‘cancer’ may elicit a slew of memories and experiences involuntarily and instigate thought processes culminating in a comorbid depression, which may radically affect how our patient perceives his or her current health and prognosis. When asked about treatment preferences, and whether the patient desires a particular course of treatment, we may have unknowingly set into action an automatic process that results in an outcome our patient might not otherwise desire.

In contrast to this proposed highly deterministic model, Baumeister and Sommer suggest that consciousness introduces explicitly indeterministic elements (Baumeister & Sonner 1997). They argue that consciousness allows us to recognize when automatic processes are occurring, and to exercise control in the behavioral process. Introducing some indeterminacy into decisional models does not contradict underdetermined decisional models, and it allows for ownership of action with accompanying ethical valence (moral praiseworthiness or blameworthiness). It reinforces the necessity of exploring the decisions we make to ensure that they are, in fact, the result of conscious mediation, and not simply the result of underlying automatic processing. I wish to stress that there are *strongly* deterministic causal factors in cognition, and that we must be aware of the myriad influences upon our choices, especially in critical situations such as forgoing treatment.

Automaticity, therefore, can be a powerful motivator for action, resulting in affective changes, goal activation, and deterministic mediators of conscious processes. These resultant changes are necessarily interactive and modifying causal elements of further cognition. As a result, we see that cognition has strongly deterministic elements at all levels of pre- and post-conscious processing. These elements necessarily conflict with our folk model of cognition, in which our decision-making is essentially free.

As such, the model that emerges from this discussion is that of a consciously mediated but often deterministic, reflexive processing in response to both external and internal stimuli which can have long term effects on affect, perception, and cognition. In short, the choices that we make can be heavily influenced, *but not necessarily determined*, by factors outside of our control. Clinicians should be very aware of the role that context and psychological stimuli have upon the decision-making process. If a patient chooses to forgo medical treatment, we would be remiss if we were not to ensure that it is done for the right reasons, and not as an automatically processed reaction to the situation in which the patient finds him or herself.

The discussion of cognition must also contain a discussion of 'mental spaces' and 'backstage cognition'—a theory of cognitive processing positing the assemblage of novel ideas and constructs from earlier ideas and constructs, occurring outside of our conscious awareness (Fauconnier 1994). Fauconnier argues that language cues give rise to cognition outside of our awareness, building complex cognitive structures that can exceed the extent of the information presented. He suggests that *any* form of thought or cognition produces such mental spaces, and stresses that these ought not to be considered simulations of reality of 'possible worlds'—consequently, we ought not to envision them as such or compare them to types of heuristics setting up simulations of possible outcomes. These elements, however, are not necessarily accessible to us consciously—we are engaging in a phenomenon called 'backstage cognition'.

The cognitive processes of which we *are* aware are surface phenomena, and merely a subset of all the phenomena occurring when we consider choices and options. Thought and judgment are much more complex processes than our everyday folk accounting would suggest, and any model of 'rational autonomy' must account for a profound empirical criticism—'rationality' isn't so rational after all. This is a very different model than what we encounter in classical models of moral agency, which posit a decision-maker as rationally mapping out the consequences of particular actions and assigning objective probabilities to each. Instead, cognition appears generally to be more *ad hoc*—judgments and meaning seem to be constructed by conceptual blending in mental spaces, rather than the results of conscious deliberation.

The material that is drawn into the blend does not have to be part of the current stimulus—it is entirely possible for one to draw upon old experiences and memories as inputs into a conceptual blend. This will be an important part of the cognitive autonomy model as well—experience and memory provide the information accessed most readily, in addition to emotional valences. We are not necessarily aware of all of the blends that our minds produce—as it is a backstage process, it is entirely possible for meanings and associations to be blended, but to be preconsciously rejected in favor of other interpretations (Fauconnier 2002). They may be rejected for a variety of preconscious reasons; while we do not presently have a full accounting of preconscious processes or reasoning (and, in light of our complexity, one might reasonably ask whether we will *ever* have such an account), we have several candidate theories in heuristics-and-biases, ecological rationality, bounded rationality, and 'fast and frugal' heuristics.

In essence, the way we think about many things is not necessarily based on the strongest information or the most accurate understanding of what information we *do* choose to focus on. Further, we are often called upon to evaluate novel situations, and in this context, we find that there are several typical constraints upon what we view as likely versus unlikely, based upon any germane or potentially relevant information we possess. We construct scenarios to evaluate how we can reach the targeted outcome; the more plausible the scenarios we discern, the more likely the target event. In principle, this serves as a common standard to decide between distinct alternatives, appreciation of the consequences of these choices, and the process culminating in the choice that maximizes the return the agent receives as measured by the common standard.

In a clinical setting, this is a description of our idealized patient and our ideal of informed consent—authentic choices predicated on an understanding of the procedures and risks involved and knowledge of the reasonably predictable outcomes. There is a problem, however—this standard is impossible. We have innate limitations on how much information we can manage in constructing these scenarios; as a consequence, we tend only to alter simple elements or factors, which may not conform to reality or may be counterintuitive (Tversky & Kahneman 1982). Further, once we construct a particular scenario, we tend to find it difficult to imagine other possibilities—we become tied or 'anchored' to one given possible explanation or course of action, which limits our ability to generate further scenarios or to see other potential outcomes. Tversky and Kahneman further note that in judging probabilities and unknowns, our decisions are only adequate if the judgment is in accord with the entire collection of beliefs held by the thinking agent. This poses a problem in assessing rationality: there is no simple way to check whether any particular set of probability judgments are compatible with the individual's collective whole. Instead, the individual simply strives for conscious and unconscious compatibility with his knowledge, assessments of probability, and his own heuristics and biases. In other terms, the individual strives to make his decision as authentic as possible.

Further modifying our knowledge pool complicates our decisional framework—we respond differently when we begin to add information into our cognitive schema. Our mind can have difficulty filtering useful information from worthless information—studies demonstrate that "people respond differently when given no evidence and when given worthless evidence. When no specific evidence is given, prior probabilities are properly utilized; when worthless evidence is given, prior probabilities are ignored" (Tversky & Kahneman 1982, 5). When information is present, we assign it decisional weight and importance, but may potentially give it undue weight, leading us to become either overly reliant upon that particular piece of information (anchoring), or overly confident in our assessment of its worth, a failure rampant across lay and professional decision makers.

Human cognition does not follow an overtly rational process like pure information processing and utility maximization; our cognition is characterized by values, emotions, prior knowledge, raw intelligence, and many other factors that do not fit nicely into this idealized model. Accounts or theories of autonomy must reflect this messiness to be sound—if our philosophy is not influenced and tempered by what we learn from neuroscience and cognitive psychology, it is an exercise not in truth but in fiction (Lakoff

1999). Special interest in concepts like backstage cognition and heuristics and biases can be traced back decades (Ashcraft 1994; Gigerenzer 1996; Gigerenzer, Czerlinski, and Martignon 2002; Gilovich & Griffin 2002; Kahneman 2011; Tversky & Kahneman, 1982), but medical ethics has not broadly integrated these findings. Models of medical autonomy from that period evidence a classical understanding of rationality and reason, and three principle models serve as examples.

Homuncular Autonomy Models

Some of these models explicitly endorse the classical cognitive model, while others only make covert appeals by linking biases and thought distortions to psychopathologies or outside influences. They propose a model of cognition which seems to suggest a high (if not total) degree of control over what influences us in our thought processes, with our only weaknesses being disease, addiction, immaturity, or dementia. The evidence of the past few decades of research in cognitive psychology and neuroscience paint a very different picture.

Veatch

The first model of note from that era comes from Robert Veatch, in which he establishes a relationship between deontological and consequentialist methodologies and principles, producing a system advocating promise keeping, beneficence, and personal autonomy (Veatch 1981). Veatch is especially concerned with liberty rights—a category of claims that prevents others from infringing upon our ability to act. Contained within this category are the right to refuse treatment and the right to control one's body. Related to this is the ability to act on the information disclosed by physicians—Veatch defends a scenario in which giving a patient unwanted information constitutes as much of an ethical violation as failing to provide information. In essence, Veatch defends autonomy rights over the provision of information, allowing for a model in which the physician must respect the autonomous decision-making of a person even in the face of obvious ignorance of salient facts. Veatch explicitly makes patient autonomy a trump—we have a duty to respect it at all costs and even in circumstances when there is sufficient justification for questioning it (e.g., psychiatric hospitalization). This need to respect autonomy extends to patients in a variety of circumstances, including those who may be experiencing terminal illnesses, which potentially impacts or compromises their ability to make decisions. Veatch recognizes that one's autonomy and moral decision-making do not exist in a vacuum—his model recognizes that the patient's moral community includes other relationships that must be factors into decision-making. This allows an outward growth of our understanding of personal autonomy, but not inward growth into our own thought processes.

Autonomy is a deontological norm in Veatch's model—it is seen as a prerequisite for evaluating a moral action (per his argument, we must address our prima facie duties before attending to their consequences). We cannot justify violating a nonconsequentialist principle, regardless of the good consequences our action may produce. This philosophy is understandable; it is entirely possible to imagine circumstances in which good consequences result from obviously immoral actions (e.g., peace produced by genocide). A moral system allowing for such an outcome is obviously suspect at best.

Despite this intuitive appeal, there are significant challenges to the autonomy concept as proposed by Veatch. Fundamentally, the picture of autonomy he proposes is built on an unrealistic cognitive model—he allows for illness to occasionally compromise a patient's competence (e.g., delirium), but he is more concerned with considering exceptions resulting from a patient's lack of information (in essence emphasizing the informed aspect of informed consent over the consent aspect). This is a clear deficit, as has been explored in the past few decades—we know much more about how the brain functions at a variety of levels of organization (from individual neurons to neural networks). We know that pathophysiology impacts our brains at the cellular and functional levels. We know that psychopharmacology and psychoneuroimmunology introduce additional factors to our unconscious thought processes. We have found any number of cognitive "rules of thumb" that creates shortcuts in decision-making that operate at levels we do not control. All of this creates a cognitive model far removed from what Veatch proposed. Obviously it is wrong to criticize a historical system based on recent findings, and much of the relevant work in cognitive psychology and neuroscience postdates Veatch's proposal. His argument, however, introduces a larger and recurring discussion in medical ethics contemporaneous to and following our insights into how we actually think.

Faden and Beauchamp

The second noncognitive model is that of Ruth Faden and Tom Beauchamp (1986), first published five years after Veatch. They also stress the essential (if not primary) importance of autonomy in medical ethics, defining it in terms of individual rights, and the obligations we have not to infringe on the ability of others to act. They include a variety of concepts under the umbrella of autonomy, including privacy, voluntary decision-making, and accepting the consequences of one's decisions. This position strongly reflects the root of autonomy—the principles of self-governance and self-direction. Faden and Beauchamp focus significantly on outside factors that can impact decision-making, especially the clinical staff (e.g., withholding information relevant to the treatment decision, not recognizing the patient's refusal of treatment, etc.). Just like Veatch, they place autonomy into a pluralistic system in which multiple values are weighed in ethical decision-making. Unlike Veatch, however, they do not give autonomy trump power—they envision circumstances in which beneficence and justice require us not to respect the patient's autonomy.

Their picture of autonomous agency does not posit a variety of strict criteria. They focus on a model of autonomy that meets our everyday understanding and experience of autonomy, in which autonomous actions are performed intentionally, with understanding, and without controlling influences (Faden & Beauchamp 1986, 238). They put understanding and freedom from control on continua—they recognize that these factors are not binary, and that individuals can experience degrees of understanding and coercion. Autonomy itself, therefore, exists on a continuum, with these variables interacting with each other. If an action is coerced, there is no degree of intentionality or understanding that can make it autonomous, just as no degree of intentionality or freedom can make an action autonomous if it is not understood.

Faden and Beauchamp developed their model of intentionality in light of both philosophy and psychology—the agent in question must have a concrete plan and act to follow up on it (instead of acting accidentally or on habituated and automatic behaviors). Their picture of psychological understanding is based on propositional reasoning and the degree to which an agent has justified beliefs about what he or she is doing. In order to demonstrate understanding, the moral agent in their model must describe both the intended action and its consequences, taking into consideration that an action may be performed with something less than complete understanding or in the presence of false beliefs. The model of controlling and coercive forces requires a separate understanding of will, voluntary action, and control—they note that an agent may fully intend and will an action even if it is influenced or controlled. An agent who is being manipulated, however, is not exercising autonomy.

As with Veatch, there are elements that are intuitive and appealing—it makes sense for us to understand the role manipulation plays in undermining our ability to act autonomously, and it makes sense to integrate concepts in psychology and philosophy in defining intentionality and understanding. However, as with Veatch, there are also compelling reasons to argue that this model is still predicated on an unrealistic cognitive agent. The historical defense provided to Veatch loses some of its weight as Faden and Beauchamp's model recognized the need to integrate psychology into autonomy and Tversky and Kahneman's *Judgment Under Uncertainty* had already been published, meaning that knowledge and insight into cognitive heuristics and biases were established enough to put forth a collection of papers for broader consumption. The larger problem, however, is that research has demonstrated a number of potential internal influences which can undermine a rational agent's thought process yet which can still yield an "autonomous" decision per this model.

Beauchamp and Childress

The third noncognitive model under consideration is, by far, the most popular methodology in contemporary medical ethics—the principlism of Tom Beauchamp and James Childress (2012). Currently in its 7th Edition, their *Principles of Biomedical Ethics* has remained highly influential in the field, and students entering clinical practice are instructed in the weighing and balancing of beneficience, nonmaleficence, justice, and autonomy. The system is rightly praised for its blending of deontological and consequentialist methodologies (similar to Veatch's blend of consequentialist and nonconsequentialist approaches), which produces a versatility and applicability in a variety of clinical contexts.

Beauchamp and Childress do not make autonomy lexically prior in their system—they recognize that there may be circumstances in which personal autonomy interests are outweighed by other, more essential claims. However, they do place significant importance on it, maintaining a framework in which autonomy must be respected and requiring significant contextual concerns to value other principles ahead of it. They understand autonomy to involve as a minimum the ability to make one's own decisions intentionally, free from outside control, and from limitations that may prevent one from making meaningful decisions (e.g., a lack of understanding). Respecting autonomy in their model requires us to recognize patients' right to hold views and opinions, the right to make choices, and to act upon their opinions and beliefs. They argue that this respect requires both positive and negative duties from us: obligations to disclose information and foster autonomous decision-making, as well as obligations to avoid imposing constraints on autonomous action. This duty does not extend to patients experiencing diminished autonomy, like immature children, those who are ignorant or cognitively incapacitated, or those who are being coerced or exploited. Thus, our obligations to those with diminished capacity to make medical decisions can be different from our obligations to an uncompromised patient.

Beauchamp and Childress tie their discussion of autonomy to competence, noting that the defining criteria of the autonomous patient and competent patient are "strikingly similar" despite having distinct meanings (*Ibid*. 116). They argue that we should not adopt global standards of competence (i.e., that we should understand judgments of competence to be task-specific) because there are significant difficulties in the validity and reliability of current tests for incompetence—the "evidence" of incompetence isn't necessarily reliable. Instead, when we are concerned about a patient's competence, we should examine her ability to understand her current circumstances and the information she has received, reason about her life decisions, and formulate a choice or preference. In light of this, they suggest that as the risk of a decision increases (for instance, the risk of death), we can reasonably ask for a greater level of *evidence* supporting a decision, but not a greater level of *competence*.

Beauchamp and Childress are not unaware of psychological issues in decision-making. They are aware of differing levels of understanding, the impact of framing effects, difficulties in processing risks, and other elements that can lead patients into false beliefs, and as a result they argue that clinicians ought to challenge patient perceptions and choices in order to better their autonomy (*Ibid*. 137). They also recognize that there are conditions that can impact the voluntariness of actions, like disease, psychiatric disorders, and drug addictions, which preclude autonomous choice and decision-making. Further, in a discussion of hard and soft paternalism, they recognize that there are cognitive biases and bounded rationality in decision-making, but they argue that these ought *not* to be bases for challenging patient autonomy, as it strays into opaque and potentially abusive hard paternalism (*Ibid*. 219). As such, they are aware of relevant challenges to the notion of a Kantian rational agent. Unfortunately, this poses significant problems for their model.

First, it suggests a contradiction, in that they encourage clinicians to challenge their patients' perceptions and choices when they are predicated on false beliefs based on misunderstanding, framing effects, and risk-processing deficits, but caution against challenging their patients' perceptions and choices when they are predicated on bounded rationality and cognitive biases, despite these factors potentially producing misunderstanding and risk-processing deficits. Second, the recognition of bounded rationality suggests awareness that there are essential limits to conscious reasoning and that there is a body of literature exploring alternative explanations for human cognition, including emotional processing, backstage cognition, dual processing models, etc. Simply put, it isn't clear how one can argue for an overly rational model of cognition when one is aware of myriad empirical data undermining this position.

The preceding analysis is not meant to fundamentally scuttle the theories discussed. They have individual strengths and weaknesses that ought to inform subsequent models. It makes eminent sense to establish prima facie duties, for instance, and to value a collaborative relationship between physician and patient. It makes eminent sense to recognize that ethics is pluralistic, and that it is unlikely that any single principle ought to carry universal and absolute weight. It makes sense to draw upon a variety of philosophical outlooks in offering justification for action, or in discerning the appropriate moral methodology for a given ethical conflict.

However, it does not make sense to predicate an ethical theory on a model of human thought that does not exist. Fiating cognitive abilities amounts to requiring us not to be human when exploring ethical dilemmas or making treatment decisions. It makes no sense to believe that we exercise control over avolitional backstage processes, or to ignore demonstrable sources of error in decision-making, especially when the choices to be made are potentially the most meaningful and most irrevocable of decisions. It makes no sense to suggest that identifiable sources of error ought not to be eliminated as much as possible, to ensure that the choice made is a genuine reflection of the patient's desires, and is not simply the disease process speaking for them. The models that follow attempt to elicit these sources of error, while reaching fundamentally different conclusions.

Cognitive models of autonomy

In contrast to the homuncular models, the cognitive models explore the backstage and automatic elements of patients making health decisions. Four principle models are examined, and the strengths and shortcomings of each are noted. A recurring theme in these critiques is that cognition is fundamentally influenced by a variety of factors not considered in the homuncular models. As such, by their very nature, they present models of autonomy that have much more empirical and ecological validity—they are autonomy models of actual human beings, rather than of idealized cognitive agents.

The first cognitive model to be considered is that of Redelmeier, Rozin, and Kahneman (1993). Contrary to the homuncular models discussed earlier, they argue that the 'ideal' decision maker—characterized by the agent who gathers all available information, calculates the risks and benefits of every option, and then selects the optimal choice—simply does not exist. Instead, actual decision-makers employ cognitive heuristics to simplify situations and find palatable solutions. Additionally, they are influenced by a variety of sources, including external and internal stimuli, and can be strongly affected by how information is framed. Minor shifts in decision context, option order, defaults, or semantics can radically alter perception and subsequent processing, and yet these are not necessarily changes of which we are aware.[4] Further, individuals can demonstrate a phenomenon called 'hindsight bias'—when individuals learn of the outcome of a given action, this knowledge affects their assessments of the likelihood of that outcome occurring. This is to say that individuals tend to ignore contradictory evidence, focus only on corroborating evidence, and overestimate the probability of the outcome. This is a significant concern in medical liability cases, for instance—arguments that a clinician "should have seen this coming" demonstrate hindsight bias. In the context of medical treatment, this can affect patients' perceptions of their current situation (e.g., 'it was inevitable that I would get cancer'), and can feed into other sources of cognitive error.

They note that many research studies fail to take into account salient features of the patient experience when exploring outcomes and efficacy. There are emotional aspects of being a patient, for instance, which are reflected in one's sense of well-being and validation. Patients, as a result, often seek medical care for sympathy and reassurance (Redelmeier, Rozin, & Kahneman 1993, 74). This presents a difficulty for research, however, in that these emotional valences and experiences are difficult to quantify in the same way as one could quantify physical or mental disability. Difficulty in measurement, however, does not translate into irrelevancy.

This emotional content complicates medicolegal issues as well. They note that the process of informed consent requires the clinician to disclose the risks, benefits, and outcomes of particular interventions. Ostensibly the patient then decides which option best suits his needs and values, but this concept does not take into account the plasticity of human emotion—his needs and values may not be the same once the intervention

[4] This really is a remarkable phenomenon. Environmental cues, for instance, have been demonstrated to be a confounding variable in research, and as such, are controlled as much as possible. Presentation order has been shown to demonstrate that individuals have a tendency to choose the last option presented to them—even if the items presented are identical—and that they will offer fabricated justifications to explain why that particular option was different than the others. The presence of defaults has also been demonstrated to affect cognition—studies have demonstrated that many individuals have a tendency simply to accept default options when presented with a choice. Finally, word choice affects salience—it has been demonstrated that individuals view information differently when it is seen as self-relevant; this perception, however, can be affected by whether the individual properly understands the terminology (e.g., there will be a difference in responses between asking someone if they are diaphoretic versus asking them if they are sweating).

has been selected and performed. They note that "psychologists have shown that people are prone to err when making decisions about long-term consequences because they fail to anticipate how their preferences will change over time" (*Ibid*. 74). This is not limited to medical settings—studies have demonstrated that attempts to forecast how one will feel produce errors in such diverse conditions as being fired from one's job to winning the lottery. We have a tendency to believe erroneously that the joy or sorrow we are experiencing now will continue unabated for the foreseeable future. As a result, they suggest that the informed consent process include an appreciation of changes over time, and that patients might benefit from including "statistics and interviews of people who underwent each therapeutic alternative months of years previously" (*Ibid*. 74). As a corollary to their suggestion, it would seem that in the case of forgoing treatment, comparable information might be included, if available.[5]

A special case is presented for patients who are experiencing a recurrence of their illness—some conditions are long-standing with periods of remission (cancer, for instance, or multiple sclerosis). Initially, one might be more inclined to accede to their wishes, as they have already experienced the positive and negative effects of the given intervention. However, even this first-hand experience is not necessarily accurate. They note that memories can also be inaccurate and subject to error.[6] As such, we should not simply defer to patients' prior experience—they may have a distorted sense of the experience (Redelmeier, Rozin, & Kahneman 1993, 74). In light of all of these concerns, they caution that the process of medical decision-making must involve clinicians providing guidance about medical information, but also about common cognitive errors. This is not, however, to claim that clinicians are in a privileged position—the clinician may employ the same kinds of errors he is seeking to prevent in his patient (Dawson & Arkes 1987).

This model provides a more accurate picture of actual cognitive processing in decision-making, but it is hardly a complete ethical theory. Rather, the article serves as an effort to translate the existing heuristic and biases literature into clinical settings, and to

[5] Clearly this may present a problem, as individuals electing to forgo treatment might not necessarily be in any shape to provide said information. Other methods of providing this information might include patient testimonials (written or video), contact with surviving family members, etc. While there are difficulties in securing this information, it is not impossible in any sense of the term.

[6] This is not a new claim—Hume, for instance, noted this phenomenon in his epistemology: our (simple and complex sense) impressions cannot be mistaken, but our recollections of those complex sense impressions are fallible. It is quite easy for us to misremember events, locations, and experiences, amplifying certain characteristics and suppressing others. As such, personal recollection and experience are not necessarily infallible guides for action.

make clinicians aware of the issues that they will have to face. More developed theories of autonomy are found in the arguments and models presented next.

Grisso and Appelbaum

Like Beauchamp and Childress, Thomas Grisso and Paul S. Appelbaum (1998) stress that the concepts of autonomy and of competence to consent to (or refuse) treatment are related, arguing that competence to consent necessarily involves four criteria. First, it is necessary that the moral agent be able to express a choice—this is not tied to any particular medium of communication (e.g., the patient does not need to be able to speak to do so), but rather, the patient must possess the ability to make his or her choices known. Second, the patient must be able to understand the information germane to the health care decision. If the patient cannot understand the information at hand, there is no way to act upon it or to voice a preference for one intervention over another. Third, the patient must appreciate the significance of the information and the expected outcomes. If there is no way for the patient to gauge risk or to weigh outcomes, there is no way for the patient to take ownership of the decision—there is a fundamental disconnect between the decision and the outcome. Fourth, the patient must be able to reason with the germane information in a manner that allows him or her to logically weigh treatment options. If a patient cannot reason and deliberate about the decision, there is no manner by which he or she can make a genuinely autonomous choice—it is akin to being asked to write a paper without having any writing implement—some organization may be possible, but clearly the ultimate goal will not be able to be realized. These four criteria are not to be understood as being 'all-or-none' principles—that is to say, each of these criteria exists on a continuum; patients manifest different abilities for each at different times. As such, like Beauchamp and Childress, Grisso and Appelbaum argue that competence is not to be understood globally, but is task specific. Ethical judgments must be cognizant of each of these criteria, but "in practice, not all of them uniformly will be 'required'" (Grisso & Appelbaum 1998, 33). Further, they reject appeals to competence criteria based popular wisdom—i.e., they reject competence criteria tied to whether most people would consider the judgment wise or correct. As such, respect for autonomy in their model requires us to respect patients' decisions despite apparent eccentricity or inadvisability (although cases of gross deficiency to make a choice do not enjoy similar protection). These criteria individually are necessary, but not sufficient, for autonomy—a marked inability to meet one of these criteria would render the autonomy of the decision suspect, but being able to meet one of these criteria is not sufficient evidence to render the autonomy of the decision beyond reproach.

The most referenced criterion is that of *Understanding*—Grisso and Appelbaum note that courts often rely upon this in decisions about competence (*Ibid*. 38). The concept, however, is quite tricky—the underlying mechanisms and processes of the 'Understanding' construct are not well known or easily defined, involving a list of physiological and psychological processes required to translate an experience into a coherent conscious model of it. This complex series of events is not the only mechanism by which cognition is influenced. There are a host of medical disorders, medications, and other injuries that can profoundly affect cognition. The ease with which disruption occurs facilitates examination and assessment—if a lack of understanding seems evident, there is reason to suspect disrupted underlying cognitive mechanisms. This is not, however, a clearly defined case of cognitive deficiency—they note that patients may *appear* to misunderstand information when the actual underlying mechanism is *miscommunication* (*Ibid*. 41).

Grisso and Appelbaum note that *Appreciation* as a competence standard refers to whether patients appreciate that they have a disorder and acknowledge the consequences of that disorder and its treatments (*Ibid*. 42–43). This use of the term parallels other authorities who refer to an absence of this appreciation and acknowledgement as demonstration of holding objectively false beliefs, explicable in terms of definite cognitive distortions. A caveat is introduced, however, in that this lack of appreciation or acknowledgement must be due to more than disagreement with the diagnosis. They note that several conditions are necessary to demonstrate that a distortion is present, rather than simple disagreement. First, the underlying beliefs the patient holds must be substantially irrational or unrealistic. There is a significant difference between doubting a diagnosis because conflicting information was presented or there is evidence of clinical disagreement and doubting a diagnosis because one believes that he has superhuman powers.[7] Their second criterion is that the belief must be the consequence of impaired

[7] A personal anecdote serves as a quick example—a patient experienced a painful swelling on her foot and lower leg following a ballet rehearsal. The first clinician to examine her in the Emergency Department ruled out torn ligaments or tendons, noting that while the swelling had abated, a rash-like discoloration remained. Operating on the premise that it was either a reaction to a bacterial or viral infection of the fascia, he contacted infectious diseases and admitted the patient for what would amount to a ten-day stay. The rash did not respond to the treatments provided, and, in fact, the antibiotics administered provoked a further reaction on the patient's hands and arms. The patient and her family became quite skeptical about the diagnosis, despite the insistence by the clinician that it was an infectious disease. Eventually

cognition or affect. This is necessary in light of the objections of established religions to specific aspects of otherwise routine treatment (e.g., Jehovah's Witnesses prohibitions on using blood products). Some of these systemic beliefs sets may be considered by the clinician to be eccentric, but that does not mean that they can be ignored. Their third criterion is that the belief must be relevant to the patient's treatment decision. If the patient is exhibiting distorted cognition that does not reflect on the treatment decision at hand, it is not germane to an assessment of Appreciation. If a patient maintains the belief that gravity does not apply to him, but manifests no treatment-relevant cognitive distortions, there is no compelling reason to doubt his ability to appreciate other information.[8]

There is a common reaction in medicine that patients are expected to react negatively to bad health news—in fact, many consider it a sign of pathology if bad news does not engender some manner of depressive reaction. However, this can have a profound impact on the course of treatment—clinicians can quite easily endorse decisions of questionable competence, as the depressive symptoms can be masked by the expected grief (Grisso & Appelbaum 1998, 51). In light of this, it may be preferable to err on the side of caution when there is evidence of cognitive distortion. Not all cases will be clear cut, and will likely require significant sensitivity to the biopsychosocial elements of the disease and its pathophysiology.

Their *Reasoning* criterion requires that patients be able to engage in logical cognitive processes using the information they understand and appreciate. As noted above, there is significant concern that one may be given information but not be able to use it. Cases of anterograde amnesia, for instance, present challenges to processing because of the speed with which information is forgotten. Alzheimer's dementia and cerebrovascular accidents near memory structures carry similar risks—they prevent individuals from

an orthopedist—a friend of the family—visited, and immediately declared that the mysterious 'rash' was simply a bruise that resulted from torn ligaments; the hospital orthopedist concurred, and the patient was discharged later that day. Clearly the patient's and family's disagreement with the diagnosis was not unreasonable or irrational. Questions about the rationality of the patient's and family's beliefs would have been more appropriately raised had she claimed that she was immune to all diseases and infections.

[8] For instance, early in my teaching career, I worked with patients with schizophrenia of a variety of severities and degrees of subsequent cognitive impairments, including auditory and visual hallucinations, perceived conspiracies and threats, irrational degrees of grandiosity, and with varying degrees of insight into their conditions. This has not prevented them from being able to engage and process information in many other areas of their lives, nor has their illness prevented many of them from being able to appreciate their clinical situation and course of treatment.

being able to work with new information presented to them. As such, clinicians assessing competence in patients with conditions similar to these ought to be aware of potential influences. Grisso and Appelbaum caution, however, that this criterion ought not be used to deny individuals their right to autonomy simply because they employ non-normative approaches to information processing. They note that most, if not all, individuals fail to meet idealized standards of decision-making in everyday situations, and that these deficits may become more apparent in times of crisis. As such, they stress that Reasoning deficits should focus on cases in which “a patient’s mental abilities are so impaired by illness or disability that even basic functioning with regard to these considerations is seriously and negatively influenced” (*Ibid*. 55).

Grisso and Appelbaum stress that certain cases merit greater attention than others—significant changes in mental functioning (generally with behavioral correlates) should serve as warning signals that cognition has been altered.[9] While refusal of treatment or evaluation may be atypical for a particular patient, that alone does not suffice to demonstrate that cognitive changes have occurred, but it should serve as a warning sign. They note that patients with organic impairments are especially prone to decisional incapacity (e.g., dementias, deliriums, etc.). They further note that while depression has been a frequently studied group, the results have varied, suggesting that the differences in the research findings may reflect different degrees of depression, with correspondingly different degrees of impairment. Further, influencing factors are additive—comorbid psychopathologies can exacerbate cognitive distortions and disabilities, which are further exacerbated by medical illness and pharmacological interventions, with polypharmacy being especially problematic (and, among elderly patients, all too common). Finally, while age itself does not *necessarily* reduce competence, they note that it does increase susceptibility to decisional impairment.

The metaphor proposed by Grisso and Appelbaum is a scale whose cups are labelled ‘autonomy’ and ‘protection’. The fulcrum is off center, allowing autonomy a natural advantage (representing social preference for personal autonomy). In the context of a patient either providing or refusing consent to a particular treatment, assessment of information is added to each side, with evidence supporting competence filling the ‘autonomy’ cup, and evidence undermining competence filling the ‘protection’ cup. Clearly in this model it requires more evidence to countermand the patient’s autonomy

[9] By this they mean patients behaving in manners contrary to their normal presentation and personality (e.g., fastidious patients who have become slovenly, gregarious patients who are withdrawn and asocial, etc.). They note that elderly patients are particularly at risk for manifesting these types of changes.

than it does to countermand the duty to protect him or her. It is very uncommon for a patient to completely lose her capacity for Understanding, Appreciation, or Reasoning—as these are continuum concepts, it is more likely that the patient's abilities will simply experience a reduced capacity. As such, clinicians need to be cognizant of the degree of impairment when balancing the metaphorical scale. The consequence of maintaining this balancing metaphor is a sliding standard of competence dependent upon risk-gain ratio analysis of the intervention in question. The fulcrum of the scale is also subject to adjustment—Grisso and Appelbaum allow the clinician to move the fulcrum dependent upon the treatment preferences of the patient. For instance, if the patient elects a procedure that has a less desirable risk-gain ratio than the intervention proposed by the clinician, the fulcrum may be adjusted slightly, requiring more evidence of competence than would normally be required. The patient, however, would need to be duly informed that greater decisional capacity must be demonstrated before the preferred treatment is initiated.

There are significant strengths in this model—for instance, its awareness of the complex interactions of illness and cognition, its understanding that normal judgment can be biased by a variety of sources not normally accounted for in other autonomy models, etc. There are some concerns, however, in that it does not acknowledge that clinicians themselves can demonstrate cognitive biases. Studies have demonstrated that clinicians can focus on one particular diagnosis and ignore others.[10] The very same cognitive heuristics that plague patient decision-makers are found in the clinical staff treating them; as such, awareness of cognitive biases and distortions is not a one-way process. The model proposed by Grisso and Appelbaum would be strengthened by a more dialogical approach, in which the distortions and biases of both physician and patient are exposed and challenged.

[10] I recall a passionate discussion I had with one psychiatrist who insisted that a patient was a chronic paranoid schizophrenic, simply because he had carried that diagnosis for several years. The difficulty, however, was that the differential was wider than this particular diagnosis—specifically, he showed considerable evidence of a frontal lobe syndrome. Specifically, he chronically abused crack cocaine (which in long-term abusers produces feelings of paranoia, as well as auditory, visual, and tactile hallucinations), per his family history he had had a traumatic brain injury prior to the onset of his symptoms, his personality was very childlike, irresponsible, and sexually preoccupied, and his affect was not flattened (flat affect is characteristic of chronic paranoid schizophrenia).

Katz

The psychodynamics of the physician-patient relationship is a key element of the autonomy model proposed by Jay Katz (2002). Katz notes that there are many definitions of autonomy, but chooses to focus on what he refers to as 'psychological autonomy'—the capacity of persons to exercise the right to self-determination, which includes their ability to reflect on the choices they have made. He further notes that current conceptions of autonomy make a significant number of psychological assumptions which go unexplored in the literature. Contemporary medical ethics is dominated by abstractions—specifically, abstract norms that generalize conduct in a manner that is inappropriate when considering how human agents actually behave. Ethicists have a tendency to rely upon the theories of Kant and Mill, among other philosophers, to relate the abstract formal norms to material situations. These abstractions contain implicit models of the human psyche which are not developed or clarified, which is unfortunate, in that "[a] careful scrutiny of many philosophical, moral, political or legal principles reveals all kinds of hidden, albeit woefully mutilated, assumptions about human nature" (Katz 2002, 108).

Paradigmatic in medical ethics are the assumptions made by Immanuel Kant—his idealized moral agent is a being of pure rationality; in the ideal agent, moral decision making will not be influenced by whims, emotions, or personal inclinations. Katz notes that current philosophers have championed this model—but the problem lies in that the model itself is untenable. Kant (1996) himself noted that he was making a distinction between an *idealized* moral agent, which he distinguishes from *actual* moral agents—it was a *theoretical* model, not a *practical* model. Kant's model recognizes only one aspect of human behavior as relevant to moral and ethical decision-making—the capacity for rational thought—but ignores or devalues many other aspects of our behavior, which is contingent upon other processes, some of which are completely irrational. Because we can be influenced by so many different aspects of our rational and irrational nature, Katz notes that Kant's model is simply impractical, and therefore is irrelevant in practical situations.

As a result, Katz adopts an autonomy radically different than Kant's ideal—psychological autonomy. Katz's clarifies his definition of the concept, noting that as an *ideal* definition, "psychological autonomy refers to the capacity of persons to reflect, choose, and act with an awareness of the internal and external influences and reasons that they would wish to accept" (2002, 111). Katz stresses that this is an ideal—the sheer volume of internal and external influences makes it impossible for a moral agent to ever be *fully* aware of them all.[11] Self-reflection and dialogic interaction with others can help to draw out unconscious influences, returning them to the control of the agent.

Katz notes that past discussion of psychological capacities of moral agents has tended to reflect psychopathology instead of underlying motives, i.e., questions of incompetence. He supports those who conclude that only the choices of clearly incompetent patients should be rejected—he argues that it quite different to recognize the sources influencing a patient and interfering with the patient's choice when one believes that they have made the 'wrong choice.' There are implicit dangers in raising psychological objections to patient autonomy—he notes that exceptions to autonomy can be too readily 'found' and that the purview of psychological objections are too far-reaching and too difficult to control. This represents a significant break between Katz's model and my own—while I can appreciate his concern regarding the ease with which questions and challenges to autonomy can be raised, it would seem that the circumstances and the choices to be made would dictate the standard of psychological evidence necessary to maintain patient autonomy (as per Grisso and Appelbaum's model). I will return to this objection below.

At this point, Katz develops the sense of the unconscious employed in his model. Employing a psychodynamic approach, he breaks from other models which suggest that unconscious elements are to be identified, evaluated, and potentially discarded. Specifically he notes the central role of the unconscious in normal decision-making—the psychodynamic perspective seeks to *understand and account for* unconscious influences, rather than *identifying and eliminating* them, as well as identifying potential conflicts between conscious and unconscious motivations. Further, the conscious/unconscious split is not the only germane factor—cognitive modelling of autonomy must take into account the rational/irrational split, as our decision-making process incorporates both. It is extraordinarily rare to find actions that stem from only one motivational source, and the rational/irrational mixture are idiosyncratic, and vary with the individual's situation. In Katz's model, 'rational' and 'irrational' reflect "capacities for adaptation to the external world, that is, persons' conscious and unconscious efforts to reconcile their internal mental processes with the external possibilities and limitations of the world in which they live. They denote persons' abilities to take reality into account and to give some

[11] In discussing internal influences, Katz is arguing from a Freudian perspective on conscious and unconscious processes, instead of the sense of the conscious, unconscious, and preconscious cognition developed here. The two are very different—the unconscious, for instance, is the domain of libidinal urges, mediated by the ego and superego in Freudian thought, while unconscious processes like heuristics and biases, information integration, and automaticity are what is meant by the term in my argument.

account of the conflicts between their inner and outer worlds to themselves and others" (*Ibid*. 117). As a result, ideal decision-making will be a dialogic process, in which the idiosyncrasies of both the patient and the clinician can be explored, leading to a greater understanding of the motivations and thought processes of both. This dialogue is not likely to reveal all unconscious motives, but it can reveal more than might be accessible solely through introspection and reflection.[12]

This model has immediate consequences for individual autonomy and liberty. Katz notes that it immediately undermines two concepts in the autonomy debate—radicalized patient autonomy, and standards of perfect understanding in the clinician. Instead, it calls for great introspection and reflection; freedom requires, in Katz's words, "constant struggle and anguish with oneself and with others" (*Ibid*. 121).

By being aware of the limits of human thought, both conscious and unconscious, rational and irrational, clinicians and patients can achieve a greater understanding and awareness of their own thoughts and motivations, and allow them to recognize how their perspectives and experience have influenced them directly and indirectly. This, in turn, gives rise to greater freedom in decision-making—the more motivational factors we are conscious of, the more control we exercise in the decision-making process. This will never produce absolute control, however, and as such, there is always an influence of unconscious and irrational factors in human thought. As such, Katz argues that the first, necessary step in self-determination is self-reflection and reflection with others. This reflection may not produce agreement with the physician and patient, but it can clear up misunderstandings and misperceptions. He still opens the door to physicians being able to interfere in patient decisions (and hence to weak paternalism in Beauchamp and Childress's sense of the term), but he stresses that neither party is asked to submit to the other, and that conversation and shared decision-making prevent significant harms.

If our aim is to facilitate autonomous decision-making, a recurring theme in multiple theories of medical ethics, it seems that conversation and mutual exploration of motives and thought processes are necessary foundational criteria. But what should be done if the patient insists on medical decisions fundamentally at odds with the opinion of the clinician? Katz argues that if we adopt the psychological autonomy model he proposes, clinicians will be required at times to accede to 'foolish choices'—as a matter of principle

[12] This is comparable to the adage that 'two heads are better than one.' Individual perception tends not to be self-challenged; the presence of another individual capable of evaluating both the situation as well as the other individuals perception.

of respect, the clinician does not possess the ability to simply overrule any decision which he feels to be ill-advised—I will address this aspect of Katz's model below.

Katz's allows for clinicians to disobey a patient's choice only when two conditions have been met (*Ibid*. 157–158). First, the consequences of the decision must pose significant risks to the patient's immediate physical condition. Katz clarifies this by limiting it to cases in which the patient's illness has interventions which have a good chance of preventing death or persistent serious injury, and when such outcomes are likely in a relatively short period of time. The second condition requires that the patient's cognitive processes are so seriously impaired that neither the clinician nor the patient can understand each other. If there is no apparent means of overcoming the communication barrier, then it is reasonable to proceed in the patient's best medical interest. These are very limited conditions, to be sure, but Katz argues that one ought to err on the side of autonomy. This does not create absolute patient autonomy, however, as Katz is cognizant of challenges which might arise as a result, and argues that if they are unable to reach an agreement, then the doctor and patient should either work within limits set by the patient or go their separate ways. As such, significant authority remains with the patient, but not total authority—respect is a principle that is not unidirectional. Many theories of medical ethics note that clinicians are not automatons—they have moral values and beliefs, just like the patient. One cannot expect a clinician to ignore her own important principles in medical decision-making.

There are significant strengths in the model proposed by Katz. It is clear that recognition of the complex cognitive processes underlying decision-making is emphasized in this model. As a corollary, recognition that both patients and clinicians carry with them their own sets of rationalities and irrationalities is an important step in shared decision-making. This model explicitly requires the identification and exploration of unconscious cognitive factors for both (or all) parties involved in decision-making, in an effort to increase understanding. This allows for critical insight that might be unavailable were one to attempt simple self-exploration and self-reflection. The emphasis on a dialogic process as a requisite first step towards self-determination clearly demonstrates the need for the patient to understand himself before he can make informed decisions. It is quite clear that we cannot make meaningful decisions if we are unclear as to what it is that we want. We can certainly make choices, but it is evident that they may not actually reflect our values or beliefs—in short, they will lack the 'self' criterion of self-determination.

However, there are some concerns about Katz's model as well. First, it is unclear that one ought to adopt a Freudian model of the unconscious, as there are significant methodological, empirical, and theoretical concerns about the Freudian model.[13] It is clear that unconscious processes influence cognition, but the empirical data and research support a model of unconscious processing quite different from Freud's theories (e.g., automaticity, heuristics and biases, and emotionally-valenced memory and recall). As such, when unconscious motivations are discussed later, it will not be in the terms Katz's proposes.

Second, the criteria set by Katz for incompetence appear to be too high. It is understandable that he would establish such strict criteria in light of the psychoanalytic model he proposes, which integrates the unconscious, but as that methodology is suspect, it seems reasonable to question the need for such restrictive criteria. This is not to say that clinicians ought to have *carte blanche* in deciding which decisions to accept or to reject, but it certainly suggests that the standards for rejecting bad choices ought to be lowered. It is clear that cognition is dependent on a variety of factors, of which we are only aware of the surface phenomena. It is likewise clear that our cognition can be affected in manners great and small at a variety of levels of reduction. It would therefore seem to be reasonable to suggest that clinicians have more leeway than Katz's proposes in challenging the decision-making process of patients, who by their nature are more vulnerable to influences due to medical illness, pharmacology, and potential psychopathology. I do not challenge the idea that patients have the right to make bad choices; I do challenge the idea that this right is an absolute, especially as the consequences of their decisions increases in severity. As suggested earlier, it seems that a quite compelling case can be made for a sliding scale of autonomy, contingent upon the severity of the predicted outcomes, with the most scrutiny applied to terminal decisions.

[13] In terms of *methodological* concerns, Freud was not research-oriented. The case studies he selected were not experiments—they were self-selected case studies designed to develop the theory, not test it. In fact, a recurrent criticism of Freudian models is that they do not translate easily—if at all—into testable variables. There are *empirical* questions as well—Freudian psychotherapy and analysis requires significant time and effort—it is common for patients to see their analyst for years before any insight is drawn. This is clearly beyond the purview of a normal in-patient stay. It is much more likely that Katz is advocating a more superficial variant of Freudian analysis, but even in this abbreviated sense, it remains unclear that the average clinician would have the requisite training or understanding needed to identify unconscious motivations. The *theoretical* concerns raised stem from Freud's own statements—as he approached the end of his life, he raised his own concerns as to whether psychoanalysis was actually helpful. If the founder of the school of thought questions its use, one ought to be skeptical about arguments built from the suspect theory.

Anderson and Lux

Higher cognitive standards are established by Anderson and Lux (2004), who argue that the keystone of autonomy and self-determination is 'accurate self-assessment,' and that autonomy is contingent upon an ability to recognize impairments in one's own cognitive capacities. They offer the clinical case of 'John'—a patient who experienced severe frontal lobe injury, which severed his optic nerves (as a result, he had no perception of light at all). As a result of his accident, John experienced a fascinating cognitive impairment—he was unaware that he was blind. Consequently, he would attempt to navigate his way around as he would were his vision normal, with the result that he would walk into walls, trip over furniture, and found himself in various dangerous situations for one who cannot see. Anderson and Lux argue that his actions ought not to be considered autonomous, not because of his visual impairment, but because of *his cognitive inability to recognize that he had a visual impairment*. This is to say, they argue, "[a]t least with respect to those actions, he was deeply alienated from himself as an agent" (Anderson & Lux 2004, 280). There are a number of types of agnosognosia (being unaware that one is unaware of a deficit)—visual, auditory, etc.—each of which pose the same kind of problem for one's self-concept. Further, there are multiple conditions which produce similar deficits in one's sense of self—V.S. Ramachandran, Oliver Sacks, and others describe neurological conditions in which a patient experiences a disconnect between sense data and association cortices, sense data and perception, perception and association cortices, sense data and emotional valence, etc.[14] Clearly it is possible to meet previously proposed criteria for autonomy and yet experience a profound deficit in self-perception. As such, it makes eminent sense for clinicians to examine self-perception for accuracy before asking patients about treatment preferences—if their self-perception is unrealistic or bizarre, there is reason to believe that decisions made upon these perceptions will also be compromised.

Anderson and Lux draw parallels to the category of 'insight into illness' in establishing their criterion of accurate self-assessment (Anderson & Lux 2004, 280). A variety of conditions manifest decreased insight—there are several psychiatric illnesses

[14] Interestingly, Ramachandran describes a procedure that temporarily alleviated post-stroke agnosognosia. Checking for nystagmus involves injecting cold water into the left ear (one of the tests performed in some brain-death protocols). Ramachandran found that individuals with a variant of agnosognosia regained an accurate picture of their physical condition (albeit temporary) following the water treatment. See S. Ramachandran and Sandra Blakeslee, *Phantoms in the Brain: Probing the Mysteries of the Human Mind* (New York: Quill, 1998) for more information.

in which the patient categorically denies any illness.[15] Inaccurate self-assessment in Anderson and Lux's sense has three criteria. First, the patient must intentionally undertake a given task. Several authors have noted that intentional action is a requisite part of autonomy and self-determination; accidental actions are not intentional, and as such, are not dependent upon an agent's belief about their skill in performing said action. The second criterion is that the agent believes that she will be able to perform the given task as it is intended. That is to say, the agent believes that she possesses the requisite skill and ability to complete the task. The third criterion is that this self-assessment of capacity must be inaccurate. Specifically, the agent objectively must not possess the requisite skill or ability in question. It must be demonstrable that the agent possesses a deficit that she does not believe she has.

When erroneous beliefs are examined, these self-perceptions are not understood in terms of whether they are subjectively reasonable, but rather whether they correspond with the facts of the case. This lack of insight does not translate into global incompetence—like Beauchamp and Childress's competence model, it is a task-specific deficit. As such, we see that clinicians assessing insight must possess an accurate understanding of the degree of skill necessary to complete the task in question—if the evaluator's criteria for normal function are set too high, it is entirely possible that competent individuals will be judged incompetent. This is not the only continuum involved in testing accurate self-assessments—in addition to standards varying with the task, the self-assessment itself is a statement of probability. Further, Anderson and Lux argue that there is no single threshold for accuracy, and hence no threshold for autonomy—for most individuals and for most occasions, a general self-assessment of one's capacities should suffice. They suggest that the cases in which inaccurate self-assessment produces non-autonomous actions will be severe enough as to be immediately recognizable (e.g., stumbling into furniture that one cannot see, but claiming no visual impairment). Some agents are able to recognize that they are experiencing cognitive deficits, and can act to correct them or to incorporate them into their cognitive modeling. They argue that the capacity (and hence the autonomy) of these individuals is still compromised in some degree, but less than it was before (maintaining the continuum

[15] For instance, I worked with a patient for several years who maintained vociferously that while he was the son of a famous martial artist, was engaged to/married to/dating a pop starlet (the relationship would change from day to day), was a commander in the Navy, Air Force, and Army, and was designing ships for NASA, all while playing with the band Metallica, he was most assuredly not schizophrenic.

approach to autonomy. They further note that just as individuals with cognitive deficits can overestimate their abilities, so too can they underestimate their abilities.[16]

Anderson and Lux stress that the establishment of non-autonomous actions requires more than simple demonstration that the patient is making poor choices or has some unjustified beliefs. They suggest that autonomy does include the ability to make mistakes. As such, they stress that in utilizing their proposed criteria, it must be clear that the deficit in question is preventing the agent from exercising self-governance—i.e., there must be something inherent in the deficit that prevents autonomy itself. There are several methods by which this may be assessed, and Anderson and Lux focus on two in particular. First, it is possible to explore the causal link between the action and the source of the action—if the action occurs in such a way as to prevent evaluation of the motives behind one's action, then the causal pathway has been disrupted, preventing the agent from taking ownership of the action. This is a key concept, and one which will be revisited later. The second method by which ownership of the action can be disrupted concerns problems in integrating the action with its motivations—the agent cannot make sense of his motives or is alienated from them (i.e., the agent experiences a baffling "Why did I do that?" moment). If the agent cannot understand and reconcile his motivations with his actions, there is reason to believe that they are non-autonomous. Anderson and Lux note that these two concerns demonstrate the need for integrated actions, as well as a means of registering that integration has not occurred—a feedback mechanism, in short. They note that this feedback mechanism "must be constituted in such a way that the unintelligibility surfaces. For to the extent to which one is unable to note the internal tensions, one is without this compass, which is so crucial for guiding one's actions in the manner we dub 'autonomous.' And this is why rigidly inaccurate self-assessments undermine autonomy" (*Ibid*. 284). In short, absent this feedback mechanism, our compass is broken, and we have no way of knowing whether we are moving in the right direction. For all we know, instead of reaching our goal, we could be simply traveling in circles. The primacy of accurate self-assessment carries with it a three-fold advantage: first, it is neutral in regards to competing theories; second, it is more plausibly linked with self-direction in autonomy; and third, it is more empirically supported in clinical neuroscience (*Ibid*. 285).

[16] In fact, this was a frequent topic in the individual and group therapy sessions held in the behavioral health hospital in which I worked. We helped our patients understand and develop their physical, occupational, and psychological skill sets and resources.

The aspect of Anderson and Lux's analysis that is most crucial to the argument developed here is that they extend it to cover mental as well as physical incapacities. Factors like automaticity, cognitive heuristics and biases, and emotional valencing occur outside of our awareness, and constitute significant but correctable sources of error and distortion. It would seem that these types of errors dovetail with Anderson and Lux's analysis; it is necessary to note, however, that they focus their analysis on traumatic brain injuries, rather than on phenomena of cognitive psychology. However, as the psychological phenomena in question have physical bases, it seems evident that such considerations as Anderson and Lux propose ought to be extended to them as well.

As with the other cognitive models proposed, there are significant strengths in Anderson and Lux's model. Meaningful self-direction is impossible if one's compass is flawed and there is no way to check it. To the extent that we can become aware of our own cognitive shortcomings, we can correspondingly increase our personal autonomy.

There are weaknesses to be found, however. First, it is unclear how far back or how deeply they are willing to extend their cognitive analysis. The kinds of deficits produced by the conditions Anderson and Lux consider also produce systematic error, since they produce a recurring mistaken belief. It is unclear, however, whether Anderson and Lux intend for their argument to be extended to the automatic and backstage elements discussed in the present argument. If they are unwilling to extend their analysis to these types of cognitive errors, it would seem a rather arbitrary distinction, and the autonomy model proposed would certainly require clarification.

The second weakness is that while the model raises compelling arguments, it does not establish a clear metric for establishing non-autonomous actions. They do specify some criteria, but they also place these criteria upon continua, which allows for significant room for interpretation. For the autonomy standard to be meaningful, it would seem that a little more structure or clarity is needed for clinical application beyond claims that distortions and corresponding non-autonomy will be immediately recognizable.

A third concern is that this is not a fully-developed theory of autonomy. To be fair, it does not seem to be intended as such, but the criterion of accuracy in self-perception is a necessary, but not sufficient, element of autonomy. It is quite clear that individuals can act in non-autonomous ways while maintaining accurate perceptions of their abilities. Additional criteria, as have been explicated in the previously discussed models, are critical to an accurate and meaningful picture of autonomy.

Conclusion

The model that emerges from this discussion must necessarily take into account multiple factors drawn from the strengths of the homuncular and cognitive models of autonomy. Four key categories of autonomy criteria emerge—foundational, medical, psychiatric, and psychosocial. Each of these categories is necessary for an autonomous action, but none are sufficient. Each will be explored in turn.

Before presenting them, however, there are several caveats. First, it must be made clear that this model ought only to be considered applicable to end-of-life decisions. It is quite clear that this kind of decisional process has little day-to-day validity—the elements discussed are not part of everyday decision-making. However, as has been suggested earlier, a compelling argument can be raised that as the consequences of our decisions become more severe, greater evidence is needed that the action is autonomous. In terminal decisions, it is unclear why a lower evidentiary standard should be preferred. Second, this model is intended for use in cases when a patient is awake, aware, and able to voice her own preferences. Last, quite obviously this should not be understood as a fully developed theory of medical ethics, nor should it be seen as anything other than criteria necessary for autonomous action as evidenced by the theoretical and empirical challenges raised to the autonomy models found in contemporary theories. It is quite possible to incorporate this understanding of autonomy in existing models (e.g., substituting a cognitive model of patient autonomy would not fundamentally undermine Beauchamp and Childress's principlism), albeit in some more than others (this model *does* present a fundamental challenge to models giving disproportionate weight to autonomy, e.g., Veatch).

Medical Criteria of Autonomy

Medical criteria concern issues that are the traditional purview of medical treatment; i.e., these are routine elements that recur in many theories of medical ethics, and are the least likely to cause concern and controversy. There are two key medical criteria for patient autonomy: the absence of a medical condition which directly affects cognition to the point of incapacity (which I will refer to as Structural Integrity), and access to the information typically required for informed consent. Both of these criteria are continuum-based, as disease processes result in different degrees of impairment, and some pieces of information might be more relevant or available than others.

Structural Integrity

The most significant challenge to patient autonomy in the models discussed is a physical impairment which prevents the patient from taking in information or processing it. Dementia, delirium, traumatic brain injury, cerebrovascular accidents, etc., can exert profound effects on the ability of the patient to take in new information, make their preferences known, form associations between concepts or words, etc., all of which are necessary elements of cognition. Clearly any illness which fundamentally disrupts this process prevents the patient from making a meaningful decision. However, because the effects of these illnesses are not uniform, it would be inappropriate to make blanket statements about the degree to which subsequent actions are autonomous or non-autonomous. As such, a threshold point would need to be established, which could employ any of a number of psychiatric and neurological tests (e.g., the Mini Mental Status Exam).

Informed Consent (or Refusal)

The standard protocol for medical intervention involves securing the informed consent of the patient. While the standards of this vary from state to state (e.g., whether the 'batting average'—the clinicians success rate with the suggested treatment—is required disclosure), there is enough commonality to require that the patient be provided with information concerning the nature and purpose of the intervention, alternative interventions (including non-intervention) and their outcomes, risks, probable outcomes of the intervention proposed, etc. This information should be presented in normal language, and should not require the patient to have extraordinary education to understand it. State standards of informed consent could suffice for threshold points (and due to variance, this criterion exists along a continuum).

Foundational Criteria of Autonomy

Foundational criteria of autonomy refer to underlying psychological structures of the decision-making process. Foundational structures are primary and fundamental—absent these criteria, significant doubt can be raised about the autonomy of the patient's decision. There are five criteria in this category: the ability to consider, make, and make known one's preferences (which I will refer to as capacity for preference); intentionality in action; accurate self-assessment; awareness of common sources of cognitive error (which I will refer to as bias vigilance); and dialogue aimed at self-discovery, which includes the willingness to participate in dialogue. There is no lexical priority for these criteria, and they fit into both absolute and continuum scales.[17] Each of these requires further exploration and clarification.

Capacity for Preference

In this criterion, the moral agent engages in reflection upon the treatment options open to her, weighs their strengths and weaknesses as she understands them, and makes her preferences known in some manner to the clinician (ideally through a contemporaneous statement). By its very nature, this will post challenges, as the interpretation the patient gives to the treatment option will be contingent upon her perception and understanding, which may require further discussion and dialogue with the clinician, to ensure as much accuracy as possible. This capacity for preference is not absolute, in that patients will differ in both the degree of their preferences as well as their ability to communicate them. Patients unable to weigh information or express preferences due to cognitive impairment or illness ought not to be considered autonomous agents, and treating clinicians should defer to a best-interest standard until the impairment is resolved or a proxy decision-maker is identified.

Intentionality

Several theories have noted the necessity of this criterion. For an action to be personally meaningful and autonomous, it must be intended and not accidental or reflexive. It is entirely possible to act without meaning to act, and a number of neurological and psychiatric conditions have demonstrated that involuntary actions can be physical or verbal. As has been discussed above, mental actions are also driven by automaticity, and therefore the agent may find herself acting or thinking in a manner she does not desire. Following earlier theories, this is an absolute scale—either one intends to act or one does not, and it is quite possible to discern between the two. Unintended actions ought not be considered autonomous.

[17] As a necessary caveat and matter of clinical significance—I realize that these proposed standards are theoretical, and may have some difficulty translating well into clinical settings (e.g., discussions of backstage cognition). This is a barrier faced by cognitive therapies in psychology, as well—the theoretical concepts will be dependent upon the underlying cognitive capacity of the patient in question. This can be resolved by using age-, understanding-, or education-appropriate terms (e.g., switching "People frequently make systematic cognitive errors in information processing." with "Sometimes we can get so used to thinking about things some way that we forget there are other ways to see it.")

Accurate Self-Assessment

Following Anderson and Lux's argument, agents must have insight into their illness. If a patient demonstrates agnosognosia, whether correctable or resistant, their autonomy has been weakened. If a patient demonstrates a consistent source of error germane to her medical decision-making process, she cannot process the information necessary to make the judgment (or can only do so in a diminished capacity), and as such lack the insight necessary to be self-directing. This analysis extends not just to awareness of physical injury, but also to persistent cognitive errors and distortions. This criterion exists along a continuum, with autonomy increasing as the degree of accurate self-assessment increases.

Bias Vigilance

Given that cognitive biases and sources of error are so prevalent in 'normal' cognition, and that special circumstances may exist in patients with depression, patients must be educated regarding common sources of cognitive error. This does not mean that the patient must hold a doctorate in psychology, but she must be made aware of the ways in which we frequently misinterpret information, emotional information, and memory. This is a continuum criteria, as patient understanding is variable. If a patient demonstrates an inability to understand backstage cognition (i.e., an inability to recognize that thought can be influenced by other conditions [environmental triggers, personal biases, heuristics, etc.]), there is reason to question her autonomy.[18] This criterion ties in directly with Dialogic Self-Discovery.

Dialogic Self-Discovery

As has been demonstrated earlier, it is quite common that we are unaware of the idiosyncratic and systematic slants we place upon the information we take in, or upon the memories we selectively recall. These biases and slants can be explored in a shared decision-making model as proposed by Katz. While the content is somewhat different than Katz's model, in that the clinician and patient are not attempting to explore the Freudian unconscious, the aim is similar—dialogic interaction can provide illumination on those processes that evade self-exploration and reflection. This criterion exists along

[18] This argument will no doubt raise significant questions, and so I feel it requires further clarification. I am not arguing that if the patient is *skeptical* about the information they are not autonomous—simple examples can demonstrate heuristical thinking, which should permit the patient to at least be willing to entertain the idea, in an effort to facilitate Dialogic Self-Discovery. If a patient demonstrates a profound *inability* to conceptualize backstage cognition, there is reason to suspect compromised autonomy.

a continuum for two reasons: first, patients will have varying degrees of insight, so the amount of benefit from dialogic interaction will vary from patient to patient; and second, patients will have varying degrees of willingness to participate in dialogic self-discovery. The more open a patient is to self-discovery, the greater the likelihood of an autonomous action resulting. If a patient categorically refuses to engage in dialogic self-discovery, there is reason to suspect compromised autonomy, but not necessarily proof.[19]

Psychiatric Criteria of Autonomy

There is only one principle psychiatric criterion of autonomy: the minimization of any psychiatric comorbidity (which I will refer to as psychiatric minimization).

Psychiatric Minimization

Given the documented underdiagnosis of depression and other depressive disorders in common medical illnesses, given the effect of depression on morbidity and mortality, and given the influence depressive disorders can exert on a patient's cognitive process, it is important to identify and account for any psychiatric comorbidities, and to attempt to minimize their effect on the patient's thought process. This may employ a trial period on an anti-depressant or mood stabilizing medication, cognitive therapy or another talk-based intervention, etc., in an effort to isolate and control thought processes stemming from a depressive disorder instead of the patient's own expressed values. This criterion exists along a continuum, as the severity of depressive disorders varies. This criterion is linked with the psychosocial criterion of authenticity.

Psychosocial Criteria of Autonomy

Psychosocial criteria of autonomy refer to the relational individual—i.e., it recognizes that the individual exists as part of a network of relationships which can exert influences—as well as referring to the narrative individual—i.e., the individual as she exists over time. There are two essential psychosocial criteria: the minimization

[19] There is also the possibility that the patient simply does not want to discuss the matter any further for a variety of reasons (e.g., irritation with the clinical staff, fatigue, pain, personality disorder, desire for privacy, guilt, crisis of faith, etc.). In the event that a patient expresses unwillingness to engage in dialogic self-discovery, it would behoove the clinical staff to identify and document the reasons for refusal, alleviate whatever conditions are immediately preventative (e.g., fatigue or pain), and attempt at a later time, when the patient may be more receptive. Reluctance or refusal are not necessarily indications of compromised autonomy.

of external coercion (which I refer to as coercive minimization) and the ownership and congruence of the individual's choices (authenticity). Both of these criteria are based on continua—recognizing that coercion and authenticity are not all-or-none principles.

Coercive Minimization

Moral agents do not exist in a vacuum—even the choice to forgo medical treatment involves at least two people (physician and patient). As such, it makes no sense to fiat a model of radical individualism, as there is significant empirical refutation of this idea. The choices that we make in life affect other individuals in a variety of ways, some strongly and others weakly. This is not unidirectional, however—the relationships in which we engage, personal and professional, influence how we approach problems and decisions. Some relationships can exert significant influence—our motives can shift from egoistic to altruistic, focusing more on how a decision affects someone else than how it affects ourselves. Further, our decisions can be manipulated by others, through bad information and deception, emotional appeals and threats, etc. Most systems of medical ethics reject such manipulations as fundamentally undermining autonomy, a position advocated here as well. This is not to attempt to argue for radical individualism, as this seems to be untenable. However, it does seem plausible that a proper accounting of personal autonomy should attempt to minimize the coercion applied to any individual—it is unlikely that *all* forms of coercion can be accounted for and prevented, but in a decision as serious as the choice to forgo medical treatment—a terminal decision—it seems clear that one would seek to minimize any *undue* influence.

Authenticity

The authenticity criterion is complicated—on the one hand, it is intuitively reasonable to desire for decisions to reflect the values and choices an individual has taken to be her own; on the other hand, humans have the capacity to change, and that inherent plasticity makes it difficult to insist that the individual act in accordance with the same principles at every point in his or her life (e.g., changing faiths from Roman Catholicism to agnosticism, or vice versa). A compromise position would seem to have individuals explore their contemporaneous values, in light of the other cognitive criteria, and in a dialogic process, in an effort to establish which principles should be considered authentic. The individual's decision could then be examined in light of the congruence between contemporaneous, reflected values and the decision made, with incongruence suggestive of compromised autonomy.

The autonomy model proposed above is no doubt open to criticism, as some claims (e.g., authenticity) have been controversial in the literature. However, they are reasonable criteria, when examined in light of the homuncular and cognitive models of autonomy discussed earlier—there is a compelling reason for each element, and the absence of any of them raises fundamental questions as to the autonomy of the action in question.

Psychology and neuroscience have demonstrated that consciousness, our day-to-day perception, our sense of self and identity, judgment, emotions, and intuitions are all predicated upon a number of causal cognitive elements that are outside our awareness—the bulk of our cognition is deterministic and preconscious. This determinism opens up avenues of undue influence into processes we normally assume to be under our control—it should be clear that this assumption is mistaken at best, inhuman and pernicious at worst. We should not abandon ourselves to blind determinism, however—we possess the ability to reflect upon our motivations, and to engage in dialogic interaction with others, who may bring aspects of ourselves to the fore which would remain otherwise inaccessible. As a result, we can take back a measure of control, but only if we engage in honest dialectic and dialogue with others.

In the context of patient autonomy and decision-making, the necessity of this dialogical process is especially evident—patients are already physically compromised, potentially in ways that can exert conscious and unconscious influence over their decision-making processes, above and beyond the normal potential sources of error found in heuristics and biases. Clinicians should be alert for such influences, recognizing that a medical illness can easily mask a deeper psychopathology. Affective disorders are very common, occur more in patients than in the general population, and tend to go unrecognized or dismissed as a normal reaction to their illness. The effect of these disorders, however, is quite pernicious. They fundamentally affect the efficacy of therapeutic interventions, morbidity and mortality, and rate of recovery—ignoring, dismissing, or failing to identify a comorbidity compromises the treatment of the obvious illness. By only treating the surface pathology, we potentially ignore the deeper wound.

Many contemporary models of autonomy suffer from similar shortcomings—while ethics seeks to inform itself of philosophical, legal, theological, and medical constructs, it all too easily ignores the psychological, an unfortunate irony in light of the fundamental connection between cognitive and clinical psychology and ethical ideals of autonomous choice. Ethical theories that dismiss or fail to address psychological constructs are groundless; models derived from inhuman absolutes are so much fancy and fiction. What good is it to describe models of cognition that have little resemblance to how we actually think?

The present autonomy model suggests that decision-making is a complex construct necessarily containing rational and emotional elements, intuitive judgments, and, as a result, potential sources of error. This seems to gel with day-to-day experience—many decisions are made by gut instinct and intuition, instead of a Cartesian rational process methodically and algorithmically exploring all possible influences, outcomes, and variables. This deterministic model gels with the phenomenon of basing day-to-day decisions upon distal causes—early education and environment, role models, learned behaviors, etc. This model suggests that as the severity of the outcomes increases to terminal, increasing reflection upon the causes and motivations of the decision is required—that a genuinely autonomous choice will explore the agent's motivations, identifying and judging the appropriateness of each influence, determining if it is congruent with the value system adopted by the agent as a whole. Decisions stemming from inauthentic elements of the self fundamentally are not expressions of autonomy; if a patient is forgoing treatment, whether to avoid suffering or actively to choose death, we would be remiss not to ensure that it is *her*, and not *her pathology* making the choice. Anything less would surrender autonomy to expediency, would surrender authenticity to apathy, and would surrender insight to obfuscation. The capacity for self-reflection appears to be a defining characteristic of being human—we would do well to use it when we face terminal choices.

References

Anderson, Joel, and Warren Lux. 2004. "Knowing your own strength: accurate self-assessment as a requirement for personal autonomy." *Philosophy, Psychiatry, and Psychology* 11 (4): 279–94.

Ashcraft, Mark H. 1994. *Human Memory and Cognition.* New York: HarperCollins College Publishers.

Bargh, John A. 1989. "Conditional Automaticity: Varieties of Automatic Influence in Social Perception and Cognition." In *Unintended Thought*, edited by James S. Uleman and John A. Bargh, 3–51. New York: Guilford Press.

Bargh, John A. 1997. "The Automaticity of Everyday Life." In *The Automaticity of Everyday Life*, 1–61. Mahwah: Lawrence Erlbaum Associates.

Bargh, John A., and Melissa J. Ferguson. 2000. "Beyond Behaviorism: On the Automaticity of Higher Mental Processes." *Psychological Bulletin* 126 (6): 925–45.

Baumeister, Roy E., and Kristin L. Sommer. 1997. "Conscious, Free Choice, and Automaticity." In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 75–81. Mahwah: Lawrence Erlbaum Associates.

Beauchamp, Tom L., and James F. Childress. 2001. *Principles of Biomedical Ethics.* 5th. New York: Oxford University Press.

—. 2012. *Principles of Biomedical Ethics.* 7th. New York: Oxford University Press.

Berkowitz, Leonard. 1997. "Some Thoughts Extending Bargh's Argument." In *The Automaticity of Everyday Life*, 83–94. Mahwah: Lawrence Erlbaum Associates.

Butkus, Matthew Allen. 2006. "Depression, Volition, and Death: The Effects of Depressive Disorders on the Autonomous Choice to Forgo Medical Treatment." Pittsburgh, PA: Duquesne University. doi:10.13140/2.1.3236.9284.

Carver, Charles S. 1997. "Associations to Automaticity." In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 95–103. Mahwah: Lawrence Erlbaum Associates.

Chapman, Gretchen B., and Eric J. Johnson. 2002. "Incorporating the Irrelevant: Anchors in Judgments of Belief and Virtue." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 120–38. New York: Cambridge University Press.

Chase, Valerie M., Ralph Hertwig, and Gerd Gigerenzer. 1998. "Visions of Rationality." *Trends in Cognitive Science* 2 (6): 206–14.

Clore, Gerald, and Timothy Ketelaar. 1997. "Minding our Emotions: On the Role of Automatic, Unconscious Affect." In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 105–20. Mahwah: Lawrence Erlbaum Associates.

Conly, Sarah. 2013. *Against Autonomy: Justifying Coercive Paternalism.* Cambridge: Cambridge University press.

Dawson, Neal V., and Hal R. Arkes. 1987. "Systematic Errors in Medical Decision Making: Judgment Limitations." *Journal of General Internal Medicine* 2: 183–187.

Donchin, Anne. 2001. "Understanding autonomy relationally: Toward a reconfiguration of bioethical principles." *Journal of Medicine and Philosophy* 26 (4): 365–86.

Eddy, David M. 1982. "Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities." In *Judgment Under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 249–67. New York: Cambridge University Press.

Einhorn, Hillel J. 1982. "Learning from Experience and Suboptimal Rules in Decision-Making." In *Judgment Under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 268–83. New York: Cambridge University Press.

Evans, Mark D., and Steven D. Hollon. 1988. "Patterns of personal and causal inference: implications for the cognitive therapy of depression." In *Cognitive Processes in Depression*, edited by Lauren B. Alloy, 344–77. New York: Guilford Press.

Faden, Ruth R., and Tom L. Beauchamp. 1986. *A History and Theory of Informed Consent.* New York: Oxford University Press.

Fauconnier, Gilles. 1994. *Mental Spaces: Aspects of Meaning Construction in Natural Language.* New York: Cambridge University Press.

Fauconnier, Gilles, and Mark Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities.* New York: Basic Books.

Gawande, Atul. 2002. *Complications: A Surgeon's Notes on an Imperfect Science.* New York: Picador.

Gigerenzer, Gerd. 1996. "On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky." *Psychological Review* 103 (3): 592–596.

Gigerenzer, Gerd, Jean Czerlinski, and Laura Martignon. 2002. "How Good Are Fast and Frugal Heuristics?" In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 559–81. New York: Cambridge.

Gilbert, Daniel T., Elizabeth C. Pinel, Timothy D. Wilson, Stephen J. Blumberg, and Thalia P. Wheatley. 2002. "Durability Bias in Affective Forecasting." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 292–312. New York: Cambridge University Press.

Gilovich, Thomas, and Dale Griffin. 2002. "Introdution—heuristics and biases: then and now." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 292–312. New York: Cambridge University Press.

Griffin, Dale, and Amos Tversky. 2002. "The Weighing of Evidence and the Determinants of Confidence." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 230–49. New York: Cambridge University Press.

Grisso, Thomas, and Paul S. Appelbaum. 1998. *Assessing Competence to Consent to Treatment: A Guide for Physicians and Other Health Professionals.* New York: Oxford University Press.

Hájíček, P. 2009. "Free will as relative freedom with a conscious component." *Consciousness and Cognition* 18: 103–109.

Homan, Richard W. 2003. "Autonomy reconfigured: incorporating the role of the unconscious." *Perspectives in Biology and Medicine* 46 (1): 96–108.

Isen, Alice M., and Gregory Andrade Diamond. 1989. "Affect and Automaticity." In *Unintended Thought*, 124–52. New York: Guilford Press.

Jennings, Bruce. 1998. "Autonomy and difference: The travels of liberalism in bioethics." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 258–69. Upper Saddle River: Prentice Hall.

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

Kant, Immanuel. 1996. *Critique of Pure Reason*. Translated by Werner S. Pluhar. Indianapolis: Hackett.

Katz, Jay. 2002. *The Silent World of Doctor and Patient.* Baltimore: The Johns Hopkins University Press.

Lakoff, George, and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought.* New York: Basic Books.

Lane, Robert E. 2000. "Moral blame and causal explanation." *Journal of Applied Philosophy* 17 (1): 45–58.

Levy, Daniel A. 2003. "Neural holism and free will." *Philosophical Psychology* 16 (2): 205–228.

Light, Donald W., and Glenn McGee. 1998. "On the embeddedness of bioethics." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 1–15. Upper Saddle River: Prentice Hall.

Logan, Gordon D. 1989. "Automaticity and Cognitive Control." In *Unintended Thought*, edited by James S. UIeman and John A. Bargh, 52–74. New York: Guilford Press.

Meynan, Gerben. 2010. "Free will and mental disorder: Exploring the relationship." *Theoretical Medicine and Bioethics* 31: 429–443.

Miller, Dale T., and Marlene M. Moretti. 1988. "The causal attributions of depressives: self-serving or self-disserving?" In *Cognitive Processes in Depression*, edited by Lauren B. Alloy, 266–88. New York: Guilford Press.

Mischel, Walter. 1997. "Was the Cognitive Revolution Just a Detour on the Road to Behaviorism?" In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 181–186. Mahwah: Lawrence Erlbaum Associates.

Müller, Sabine, and Henrik Walter. 2010. "Reviewing Autonomy: Implications of the Neurosciences and the Free Will Debate for the Princple of Respect for the Patient's Autonomy." *Cambridge Quarterly of Healthcare Ethics* 19: 205–217.

Nisbett, Richard E., Eugene Borgida, Rick Crandall, and Harvey Reed. 1982. "Popular Induction: Information is not Necessarily Informative." In *Judgment Under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 101–16. New York: Cambridge University Press.

Parks, Jennifer. 1998. "A contextualized approach to patient autonomy within the therapeutic relationship." *Journal of Medical Humanities* 19 (4): 299–311.

Redelmeier, Donald A., Paul Rozin, and Daniel Kahneman. 1993. "Understanding patients' decision: cognitive and emotional perspectives." *Journal of the American Medical Association* 270 (1): 72–76.

Roessler, Beate. 2002. "Problems with autonomy." *Hypatia* 17 (4): 143–62.

Schwarz, Norbert. 2002. "Feelings as Information: Moods Influence Judgments and Processing Strategies." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 534–547. New York: Cambridge University Press.

Schwarz, Norbert, and Leigh Ann Vaughn. 2002. "The Availability Heuristic Revisited: Ease of Recall and Content of Recall as Distinct Sources of Information." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 103–19. New York: Cambridge University Press.

Shepherd, Joshua. 2012. "Free will and consciousness: Experimental studies." *Consciousness and Cognition* 21: 915–927.

Simon, Herbert A. 1990. "Alternative Visions of Rationality." In *Rationality in Action: Contemporary Approaches*, edited by Paul K. Moser, 189–204. New York: Cambridge University Press.

Sloman, Steven A. 2002. "Two Systems of Reasoning." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 379–96. New York: Cambridge University Press.

Slovic, Paul, Melissa Finucane, Ellen Peters, and Donald G. MacGregor. 2002. "The Affect Heuristic." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 395–420. New York: Cambridge University Press.

Smith, Eliot R. 1997. "Preconscious Automaticity in a Modular Connectionist System." In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 187–202. Mahwah: Lawrence Erlbaum Associates.

Stanovich, Keith E., and Richard West. 2002. "Individual Differences in Reasoning: Implications for the Rationality Debate." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 421–40. New York: Cambridge.

Tait, Rosemary, and Roxanne Cohen Silver. 1989. "Coming to Terms with Major Negative Life Events." In *Unintended Thought*, edited by James Uleman and John A. Bargh, 351–82. New York: Guilford Press.

Tauber, Alfred I. 2003. "Sick autonomy." *Perspectives in Biology and Medicine* 46 (4): 484–95.

Turner, Mark. 2000. "Backstage Cognition in Reason and Choice." In *Elements of Reason: Cognition, Choice, and the Bounds of Rationality*, 264–86. New York: Cambridge University Press.

Tversky, Amos, and Daniel Kahneman. 1982. "Availability: A Heuristic for Judging Frequency and Probability." In *Judgment Under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 163–178. New York: Cambridge University Press.

Veatch, Robert M. 1981. *A Theory of Medical Ethics.* New York: Basic Books.

Weinstein, Neil D. 2003. "Exploring the links between risk perceptions and preventive health behavior." In *Social Psychological Foundations of Health and Illness*, edited by J. Suls and K. A. Wallston, 22–53. Malden: Blackwell Publishing.

Wolpe, Paul R. 1998. "The triumph of autonomy in American bioethics: a sociological view." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 38–59. Upper Saddle River: Prentice Hall.