

Ep #72: The Hidden Dangers of Overcapacity



Full Episode Transcript

With Your Host

John E. Grant

[The Agile Attorney](#) with John E. Grant

Ep #72: The Hidden Dangers of Overcapacity

Picture this. It's a three-day holiday weekend. Beautiful sunny weather. Your friends are out in the world living their #BestLife. And you're sitting at your desk because, quote, you were assigned more work than there are hours in the day, end quote. Sound familiar? This is one of two insidious side effects I'm going to talk about today that come from operating a law practice beyond its functional capacity. And how taking on more work than you and your team can comfortably handle is secretly sabotaging your practice, even when you think you're doing the right thing.

You're listening to the *Agile Attorney Podcast*, powered by Agile Attorney Consulting. I'm John Grant, and I help legal professionals of all kinds build practices that are profitable, sustainable, and scalable for themselves and the communities they serve. Ready to become a more Agile Attorney? Let's go.

Hey everyone, welcome back. So, this week I'm going to revisit one of the core concepts, and really a couple of core concepts, from how I approach thinking about improving the flow of work in legal systems. And actually, this is true for any system, but obviously, I care most about legal. And those are the twin concepts of, number one, the honest reckoning with capacity, and then number two, the other side of its coin, which I call the brutal assessment of priorities.

And I did a deep dive on these concepts all the way back in episode 2 of this podcast, and I definitely encourage you to go back and check that out. But I'm going to hit on some of the high points and then talk about some new things that have come to mind for me that have brought these ideas back to the forefront.

One of them had to do with a Reddit post that I saw over Memorial Day weekend, and as you've heard me say before, I read Reddit for fun and punishment, and I pay particular attention to a lot of the lawyer and legal subreddits. The other one actually comes from some work I've been doing with a client where I had one of those moments doing an on-site workshop

Ep #72: The Hidden Dangers of Overcapacity

a few weeks ago where we came to a realization about the firm being over capacity and how it was leading to some really frustrating behaviors that was driving the law firm owner a little bit crazy. So, I'll tease those for now and get back to them in a few minutes. But first, let me go back over the core concepts around the honest reckoning with capacity.

Now, the first thing I'm going to say should not be that controversial, but it bears repeating because there are some very real human cognitive biases that cut against this truth. And the truth is that your capacity is finite. And again, that's not earth-shattering news, but there's this thing called the optimism bias, which is documented in lots of behavioral science research. It tends to make us think two things are true.

Number one, that we're going to be capable of accomplishing more work than we can actually accomplish. And number two, that a particular set of tasks or a particular piece of work is going to turn out to be easier to accomplish than it actually is. And as you can imagine, that combination of things really leads to some problems. So, as my grandmother used to say, your eyes are bigger than your stomach. We think we can put more on our plate than we can actually handle in a reasonable amount of time, and that leads to this problem of overload, overwhelm, too busy, too much work, all the things that seem to be pretty normal in the day-to-day practice of law, but are not ideal and are avoidable if you take the time to be intentional about some specific steps.

And obviously, if you've been listening to my last few episodes, this concept of capacity comes up a lot for me, right? A couple of weeks ago, when I talked to RJon Robins about profitability, the concept of capacity came up in the context of once you realize which activities are more profitable in your practice, you tend to prioritize those activities, accepting them into your finite capacity, and you deprioritize other types of work that are not going to be as profitable or not going to give you the return on investment.

Ep #72: The Hidden Dangers of Overcapacity

And then last week, in my talk with Dr. Colin James about burnout and stress and trauma-informed lawyering, we talked about how capacity isn't just the time on your calendar, it is the dynamics around technical ability, and yes, time and attention, but also, and the important part about that episode was this notion of emotional capacity. And how we can run into trouble as practitioners if we're overloading ourselves in any of those aspects of our capacity.

Now, the other thing that comes into play when we're talking about capacity is your utilization of that capacity. And I talk sometimes about this concept called Little's Law, which is related to a similar concept called the Kingman's formula, and I talk about both of those in episode 5 of the podcast. But the metaphor I like to use is imagine you are juggling balls. And if you've got one ball that you are, quote unquote, juggling, of course, you're not actually juggling because you get to spend all of your time and attention, your full capacity on just that one ball, that one project or matter or task set. So you can be really, really effective in delivering that one thing.

As soon as you add a second ball to the mix, now you've got this challenge of shifting, of context switching. You're basically creating this little gap, even though you can physically hold both balls, assuming you're able to use both of your hands. Your head and your brain is still shifting back and forth, and that little gap between the balls adds administrative overhead. This is sort of a high-level management cost that is necessary for the conducting of your overall business, but it isn't directly adding value to the work on either one of those two balls that you're working on.

Now, the problem is, as you add more and more balls and start throwing them into the air, those gaps start to multiply, and they add up, and you get to this point where once you're juggling 15, 20, 50 different matters in your practice, you're actually spending a significant part of your finite capacity just on the overhead of managing all of that work. You are context switching, you are organizing, you are communicating, and those are valuable activities. I don't want to diminish the importance of doing those

Ep #72: The Hidden Dangers of Overcapacity

things, but what it means is you're spending more and more time on the gaps between the balls and less and less time on the activities needed to progress the balls themselves.

And you know, there is so much advice and there are so many tools that are designed to help you make that administrative overhead go easier, faster, smoother, whatever. But you can't get rid of it entirely. And the alternative is to be much more intentional about your capacity and to make sure that you are taking on the right amount of work relative to the people, the processes, and tools that you have that make up the capacity inside of your system.

Another way that I talk about this utilization challenge is with a freeway metaphor or a highway metaphor, right? And if you can imagine a multi-lane highway that is operating at 40, 50, maybe even 60% capacity, there's still a lot of cars on that freeway, but those cars are probably going pretty close to the speed limit.

Once you get up to about 70 or 80% capacity though, now the road is covered with so many different cars that the drivers are having to spend more of their time tracking each other, tracking the other cars on the road, looking out for dangers, looking out for obstacles. And the natural thing that happens when they're doing that is they slow down. They're slowing down for safety, they're slowing down for lots of reasons. But the reality is that a freeway operating at 70 or 80% capacity is going to be moving much, much slower than that freeway that's moving at 50 or 60% capacity.

And then, of course, a freeway that's operating at 90 or 95 or 100% capacity is slowed to a crawl, right? That is stop-and-go traffic or dead-stop traffic, and you're not going anywhere. So, even though you have got a lot of cars on that freeway, that freeway is really well utilized, that doesn't mean that those cars, or if we're talking about cases in your system, they're not going anywhere. They're very likely to be stuck.

Ep #72: The Hidden Dangers of Overcapacity

So, one of the reasons to have that honest reckoning with capacity is to make sure that you don't get yourself up to that 90 plus percent level of utilization. In fact, ideally, we want to be working closer to that 70 or 80% or even a lower level. And I should note here, when I'm talking about capacity, I'm talking about a number that you can measure. And it's something that you're going to have to dial in over time, but you know, I'm a big fan of my friend Melissa Shanahan's formulation, right? We want to work off of facts, not feelings.

And while there is definitely an emotional subjective component to capacity, we also want to talk about coming up with what is the maximum case count that we can handle, or what is the ideal case count that we should be handling, and how is that distributed across the different phases of our workflow so that we can really reflect back and we can have those feedback loops that let us know when we've got too much, too little, just right, whatever it is, we can be smarter about our decision-making.

There's also this very real, but frankly, counterintuitive phenomenon, which is if you keep fewer work items inside of your system in progress at any given moment, you will deliver each one of those work items faster. And when you can deliver work items faster, then over a longer stretch of time, you will actually deliver more work in that longer period of time by operating at a lower utilization rate than you will if you fill things all the way up to the maximum and the work just keeps getting stalled and stuck along the way.

It's the reason we meter the on-ramps on the freeway. And it's easy to think when you're stuck behind that on-ramp meter that, oh, the reason why I'm stuck here is because the system is giving preference to those cars that are already on the freeway so that they can get home or get to work or whatever faster. And because I was a little bit later, I'm being punished behind this stop sign, and I don't get to get onto the freeway sooner. And there's an element of that. But the other thing about those freeway meters is they actually are helping you get to your destination faster as well because they're helping prevent the freeway from going over capacity and

Ep #72: The Hidden Dangers of Overcapacity

getting to that gridlock stage so that once you get into the resource, you're likely to complete your journey faster. And that's true of your caseload as well.

The other pitfall of going over capacity, right? There's the administrative overhead that is planned overhead, right? You know you're going to take it on when you take on a new case. But there's this other thing that comes up that I talk about sometimes that in the lean and agile world, we call failure demand. And failure demand comes from a situation where the work isn't moving. It wasn't going according to plan to begin with, and that causes all sorts of other things to have to happen in order to track the work or communicate about the work or keep it on track.

And the most obvious manifestation of that in legal work is the client calling saying, "Where's my stuff?" Right? We hear this all the time. And if you have too much work, that work is getting stuck, work is getting lost, it's falling through the cracks, it's falling to the bottom of the priority list. And then all of a sudden the client calls and you have to respond to that. There's two things that aren't great about that.

Number one, all that time you take responding to the client isn't doing anything to move the matter forward, right? It is complete waste from an effort standpoint, and it's waste that could have been avoided if you or the firm were managing capacity properly to begin with. Number two is then you get into this squeaky wheel syndrome where you're constantly responding to the loudest client, and not only does that penalize the people that tend not to be the squeaky wheels, who tend to be the kinds of clients you like to begin with, it puts you in this ultra-responsive firefighting mode that makes it harder and harder and harder to actually feel like you can get out ahead of things. And that's a problem.

So, really quick, the flip side, as I talked about, of the honest reckoning with capacity is the brutal assessment of priorities. And this, again, is more useful when you have an actual number or at least a range of numbers to

Ep #72: The Hidden Dangers of Overcapacity

work with. But the idea is, is once you recognize that your capacity is finite and you put some sort of a quantity on that capacity, it forces you to do the next obvious thing, which is decide which things qualify for taking up that finite capacity.

And the reason that I call it the brutal assessment of priorities is that the act of prioritizing one thing is inherently the act of deprioritizing everything else. It means you have to say no to something or not now to something. And we're not great at that. We don't love that. It doesn't always feel good to do that. Although I will venture that if you get in the habit of doing this, it will feel better and better and better as you get used to it.

Okay, so one last metaphor before I get into the new stuff, which is if you're running a practice that is already at or over capacity, or frankly, if it's even close to that maximum capacity, because we don't want to be running at that 90 to 100% level, then there's this idea that I use all the time, which is if your bathtub is overflowing with water, then yes, we need to do some work to clear the drain, but the first thing we have to do is turn off the tap, or at least turn it way down.

And so, when your law practice is in a place where you're already at or close to full capacity, you really need to either stop or slow down your intake of new work. You've got to close the in-door so that you can worry about processing things so that they can make their way out of the out-door and free up that capacity to allow you to be smoother, more consistent, more predictable, more proactive in your work.

Okay, so let me tell you about this Reddit post, and it's like, I feel like I'm reading a comic strip to you. So, hopefully, this comes across well enough. But it really caught my eye. Number one, because it was a little bit sad. Number two, because it was such a clear cry for help. And number three, because I think it's a situation that so many lawyers find themselves in. And I really hope that you can avoid it.

Ep #72: The Hidden Dangers of Overcapacity

So, the title of this post is, quote, "Working Memorial Day weekend when it's beautiful weather... Tell me I'm not alone!" Exclamation point. This is an impassioned plea, right? It is a call for help, a call for compassion, a call for consortium. They want to feel like this is a normal thing or that they shouldn't feel too bad for giving up this beautiful long weekend that they frankly should be able to enjoy in the way that they want to.

And so here's the body of the post. Quote, I'm not sure what I'm looking for. I guess I want to know that I'm not alone in my misery. But I work for a really great firm that's paying me well, and I'm overall pretty happy. But the weekend work expectation is out of hand. Right now, it's a three-day weekend. It's 80 degrees and sunny out, and I'm sitting at my desk. I have been all weekend, and I'm planning to do the same thing on my day off tomorrow. I have no choice. Things are due on Tuesday. Friends are out barbecuing, and here I am. This sucks. Please tell me it's worth it and that I'm just being a baby.

Then next paragraph, I'm a second-year associate, so maybe this is something I just have to get used to, but I'm sad. Tell me I'm not alone and that it's all worth it.

All right. So that alone is a lot. And I think you're probably having some thoughts and some feelings just listening to that post. I'd like you to hold those for a minute, and then I'm going to reflect on my thoughts around it, and let's see if we agree or maybe if I alter how you think about it a little bit. As we go down the post, the most popular reply was someone who basically said what I think is the right thing. So, they said, quote, "The work will always be there. It never ends. You need to enjoy your holiday weekend. Focus on hitting the hard deadlines and getting those things done early, but don't make a habit of working every weekend or every holiday, or else you'll burn out. No one at your firm is going to tell you have to take it easy. You have to set your own boundaries."

Ep #72: The Hidden Dangers of Overcapacity

So, I think that's pretty good advice. Although I would argue that someone at the firm should be telling everyone to set the boundaries or helping to set boundaries for the firm overall. But in general, I think that comment is right. Then the original poster comes back and says, quote, "Totally. But unfortunately, these are hard deadlines. I was simply assigned more work than there are hours in the day. I'm going to try my best, but if it's not done, it's not done. I'm not sure what else to do." End quote.

And this is the thing that really jumps out at me, right? That quote, "I was assigned more work than there are hours in the day." This is the clear signal that the owners, the managers, the supervisors, whoever this person is working for inside of the firm, is not engaged in any sort of an honest reckoning with capacity. Or if they are, they simply don't care. And that's frankly terrible from my point of view. All right. So now I'm going to give you what my advice back to this associate was.

And frankly, for those of you who are workers in a law firm, this is the advice I would give you in a similar situation. And for those of you who are owners of a law firm, you need to know that this is the advice that I would give the people in your firm. And I think this is the correct advice. And if you don't like this advice, then you probably need to do a little self-reflection about your own relationship with capacity and your relationship with overwork because it's not healthy for you, it's not healthy for your team, and it's not healthy for your business overall.

So, here's my response. Quote. Number one, recognize that your capacity is finite. If you use a Kanban system or other project management tool to track your assignments, both the in-progress ones and the upcoming ones, you'll be able to know when you're reaching capacity, and that will help you know when to either say no or not right now when someone tries to give you more work. But both of those should be perfectly reasonable answers inside of a law practice.

Ep #72: The Hidden Dangers of Overcapacity

As an employee, it is not your job to manage your firm's overall capacity. As long as you're giving the proper signals when you're at or near or over capacity, then if the partners choose to disregard those signals and take in new work anyway, then they should be the ones skipping gorgeous holiday weekends. Fair warning, many, many partners will hate this because they didn't realize early enough in their own careers that they could or should be doing the same thing. We need to normalize healthy boundaries in this profession. And yes, there will be situations where unexpected circumstances will mean working late nights and weekends. That is part of the practice of law. But normal and predictable deadlines should never be one of those circumstances.

Let's face it. You are working this weekend because your firm can't or chooses not to effectively manage its capacity. And the only way to change that is to set healthy boundaries and start saying no to things. And that was the end. So, I stand by that answer. It's kind of uncomfortable, but as law firm owners, as managers, as team leads, we certainly want to have the kind of people on our team who are willing to put in that extra work. But the fact that they're willing to put in the work is something we should really cherish and protect and only use it when we absolutely positively have to. And we should not normalize working overtime, working over capacity so that we can just bring more work into an already overburdened system.

As I talk about at the top of every episode, right? I want to help you build a practice that is profitable, sustainable, and scalable. And it's that sustainability piece that we're talking about when we talk about operating within our actual capacity.

So, this other situation that came up, right? This other hazard of going over capacity is one that really hadn't occurred to me before I had this moment inside of a workshop with one of my clients. And the whole team was there. And this is a client that continues to struggle with capacity challenges. And it's one of those practices where demand is just off the charts, right? There are not enough lawyers in their area of practice to serve the community.

Ep #72: The Hidden Dangers of Overcapacity

And so, they've got their phone ringing off the hook. They feel like they need to be doing more and more of this work because it's frankly part of their mission. It also is revenue potential or opportunity, right? It's hard to say no to work when it's literally knocking on your door.

But the frustration that the law firm owner keeps having is that certain cases are just languishing. They're stalled, they're stuck, and the firm owner feels like they have to be the one sort of constantly scanning their caseload. In this case, going across their Kanban board, which makes it easier, but still they keep running into situations where something that seems like it should have been done weeks, if not months ago, is just sitting there kind of waiting for someone's attention. I've got some clients that have a metric for this, which they call WNR or Waiting For No Reason. I think that actually might be a metric from Larry Port and Dave Maxfield's book, *The Lean Law Firm*.

Now, like I said, this is a firm that's using a Kanban board, and I won't give you the whole thing, but this is the idea that you've got columns on the board, and then cards that represent matters or cases in your system, and the columns represent different phases of work. And the idea is to get the cards to flow through the phases of work as smoothly and predictably as possible, and hopefully getting them all the way to done or at least to a natural resting state.

And this particular team is using the Kanban cards and also the task lists inside of those cards to basically manage all of the work of the firm. And people are assigned tasks, and those tasks are associated with different matters, which are represented by cards on the board. And when you get all of the tasks done for a particular workflow stage, then the card gets to flow to the next stage on the board.

Here's the problem. If a person on your team has so many cards assigned to them or so many tasks inside of those cards assigned to them that, like that Reddit post said, there are more assignments on their plate than there

Ep #72: The Hidden Dangers of Overcapacity

are hours in the day, you've now done an unintentional thing, which is given them license to cherry-pick the work that they want to do instead of focusing on the work that needs to get done. Now, let me back up and talk about one other concept that I've talked about before, which is the importance of adopting a first-in, first-out policy as your default policy for how work should flow through your practice. And first-in, first-out means that you should be working on the oldest cases in your practice in order to try to get them to done before you work on the newer cases.

And the importance here is it's a default policy. It's not always going to be applicable. Work might get stalled for any number of reasons. It's out with a court, it's out with a third party, it's stuck with a client. But we still, as much as we can, want to be pushing the oldest work to move so that we can get it across that finish line. Once that gets across the done line, we do three things that are really important, right? Number one, we have delivered the final bit of value to the client. Ideally, then number two, we get paid for that work. But from a workflow standpoint, the most important thing is we've now freed up the capacity that case was taking up inside of our practice, and we can reallocate it back to the newer work.

But when you've got too much work in your system and your individual team members have more assignments than they have capacity to deliver in a particular time period, they get to cherry-pick. They get to choose from this universe of assignments and say, "Well, I'm going to work on this one that looks most interesting or easiest or otherwise attractive to me for whatever reason, and I'm going to avoid the harder ones or the less fun or the stickier cases." And because they've got all these assignments, they're effectively all important, or else they wouldn't be assigned to them. And so you've got this world where the worker is choosing what the policy is, and they're going on feel. They're not following the firm policy of moving the oldest work through the system.

And I should say, this is a problem for this particular firm even though they're using a Kanban system and even though we've tried to put some

Ep #72: The Hidden Dangers of Overcapacity

things in place like work in progress limits and other things to manage capacity, the firm still isn't following them as well as they should, and they're letting more work into the system, and it leads to this problem. But if you're not using a Kanban-based system, it can be even worse. And I've seen with so many law practice management systems, and I won't name them, but they tend to be really focused on assigning tasks to people, right?

These tools are, in my opinion, really task-happy. They let you do a lot of definition of these tasks up front, and then almost worse than that, there's a lot of them have these automations that when a matter hurts a certain point or when a date passes, the system just automatically assigns a bunch of tasks to people on your team with no regard for how much work is already on their plate.

And so these task-happy and automation-happy systems wind up doing the thing that this associate in the Reddit post is complaining about, is it just pushes and pushes and pushes a bunch of work onto that resource without acknowledging whether or not they are at or over or close to capacity. And that puts them in a personal place of overload, and then it also, like I said, gives them this license to pick and choose the kind of work that they want to do. It also adds to the administrative overhead problem I talked about earlier, right? They've got lots and lots of potential balls that they could be working on, which means that they have to scan across this universe of work and make judgment calls and think about which thing should I actually be doing. None of that work is actually moving one of your matters forward. It's simply cost, right? It's not helping deliver the legal work for the client.

So, the answer, with a good system, and there's a couple of Kanban tools that I recommend. One of them is Kanban Zone, the other is Business Map. You can read about both of them in my website. But they give you tools both at an individual worker level to sort of assess and manage their own assignments and to start to give signals back to the system and the team when they're approaching their capacity, but then also as a manager

Ep #72: The Hidden Dangers of Overcapacity

to be able to look and see, oh, what's already on this person's plate, or who on my team maybe has a little bit of capacity so that I can assign new work to them instead of to this other busy person.

Or if everyone on my team is already busy, then maybe I need to say no to some things, or maybe at least not now. And that sort of gets back to the dental chair theory of lawyering that I talked about back in episode 64.

All right. So just to wrap this up, right? Two new reasons to engage in, number one, the honest reckoning with capacity, and number two, the flip side, which is the brutal assessment of priorities. First, we want to avoid putting our people over capacity in a way that is forcing them to over-allocate on work and under-allocate on other important aspects of their lives. We want to avoid the stress and the burnout and the potential for traumatic response or emotional overload that can come from being over capacity. And like I said, it is not a sustainable practice for anyone on your team to be doing that on a regular basis.

The second thing is that as a law firm owner or manager or team lead, managing your and your team's capacity properly helps you avoid that crazy frustration of going, "Why is this thing not getting done? Why is this particular matter stuck?" And one of the reasons it's getting stuck is you've got too much overall work in your system, right? It is a freeway that has too many cars on it. And when you've put yourself over capacity, that's giving members of your team license to cherry-pick the work they want to work on and avoid the work that they're less interested in or less excited by.

So, by limiting work in process, which capacity management is a form of a WIP limit, by making sure that we've got this at least default first-in, first-out policy, that we're encouraging work to make sure that it flows through the system, that people only have a limited universe of things they can choose work from. Or frankly, what we're doing is taking some of the choosing out of the system and just saying, "Look, this is the matter that needs work. So we've got to buckle down and do that work because we've got to get it to

Ep #72: The Hidden Dangers of Overcapacity

the next phase." And if you don't like it, well, at least you're going to get through it and it's going to be done.

And what that's going to do is really encourage the work to flow. It's going to help get the cases to done more quickly, more predictably. And that is going to increase your overall throughput, which is going to increase the number of cases that you complete in a given period of time. And the counterintuitive thing is that by starting less, you will finish more. And by finishing more, you're going to get more work done in that 3, 6, 12-month period of time, and it also is going to translate to better revenue and better profitability for your firm. Revenue because you're getting more work done, and profitability because you're spending less time on administrative overhead stuff and more time on the actual delivery work.

All right, that's it for this week. I'm actually going to continue this theme of capacity next week, and I'm going to talk about some of the pitfalls that AI seem to be introducing when it comes to capacity management inside of a law practice. So, you'll want to stay tuned for that.

If you have any questions about this topic or if you want to discuss ways to better assess your own capacity inside of your law practice, or if you just want to have something that you want to hear me talk about on the podcast, please don't hesitate to reach out to me. I'm at john.grant@agileattorney.com, or you can join the Agile Attorney Community and get some information in there.

As always, I am grateful for the production support I get from the great team at Digital Freedom Productions, and our theme music is Hello by Lunar. Thanks for listening, and I'll catch you again next week.