Full Episode Transcript

With Your Host

John E. Grant

Ep #61: Delivering Your Legal Work: Direct Flight or Long Layover?

There is a lot of excitement in the legal world right now around how AI and other technology can speed up your legal delivery engine. Faster research, faster analysis, faster drafting, and I'm here for it, they all will play an important part. But here's the problem. A faster engine does not by itself create a faster race car. In fact, if that faster engine just overpowers the other components of the car or the crew that's supporting it, it can lead to some real problems.

In today's episode, I'm gonna talk about some ways to bring your law practice up to speed in ways that are focused, balanced, and designed to get your deliverables all the way across the finish line, not just record a hot lap partway through the race.

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 to the podcast. So I know that to some extent you all know this, but to kick off this episode there's something I want to call out specifically. When a client hires you to handle a legal problem for them, they of course are hoping to get the best possible outcome to that problem, and they're relying on your skill and experience to help get them there. But all things being equal, the law practice that delivers a quality outcome faster is going to generate a better client experience and ultimately command a better position in the marketplace than a firm that gets the same outcome over a longer stretch of time.

And this is obvious if you have the kind of practice that's focused on helping clients seize an opportunity, stuff like business deals or IP protection. Those clients understand the time value of money, and they definitely want to get things moving as fast as possible. But most lawyers

are helping clients work through pain points. They've found themselves facing some form of risk or navigating a complex situation. Often both.

And the reality is that most legal problems are inherently stressful, even the opportunity-seeking ones. So your clients aren't just seeking a favorable outcome. They're also often desperate to escape the uncertainty, the financial pressure, and the emotional toll that comes from living with unresolved legal issues. And you've probably seen this, right? While many clients start out focused on winning their case, the longer the matter drags on, the more likely they are to shift to, I just want this over with. Delays in the matter increase frustration, they erode trust, and they can lead to the client settling for less than they deserve, less than you know you're capable of getting for them, just to make the stress of the matter go away.

On the flip side, faster delivery doesn't just improve client satisfaction, it can change their entire experience of working with your firm. A client who feels like their matter is making clear and consistent progress towards a resolution are far more likely to stay engaged with their batter. And that steady progress builds confidence, it reduces stress, and ultimately enhances the client perception of the value of your work.

And of course, the stress from delays isn't just a client issue. Those delays create operational drag on your practice. Work getting stuck in your system means stalled cash flow, it means increased workload stress, and more opportunities for things in the matter to change, which then causes failure demand and rework.

So enter generative AI and all the promises it holds for speeding up workflows, right? All the stuff Damien Riehl talked about in last week's episode. And the stuff works. As I said in the conclusion to that episode, I absolutely believe that lawyers should be exploring and using AI and other technology tools to speed up the parts of their practice. But just speeding

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

up the parts isn't by itself enough to speed up the delivery of client outcomes.

So today I want to emphasize the importance of speeding up the delivery of legal work. And that's the key word, delivery. Because the benefit of your work doesn't come just because you've done some work. It comes when that work delivers an outcome. And the delivery is what gives it value, not the effort.

And I want to be clear, I'm not talking about cutting corners or doing anything in the name of speed that would sacrifice the sufficiency and quality of the deliverable. This isn't about rushing to a conclusion, it's about eliminating unnecessary delays in your workflow. Although, I'm going to come back to that notion of sufficiency in a minute.

Think back to that race car analogy from the top of the episode. It's not just a bigger, better engine that wins the race. It's the coordination of the entire system.

The engine, the wheels, the body, the fuel, the driver, the pit crew, even the track conditions in the other races. All of the parts need to work together to navigate the course and deliver victory. So today I want to break down the two primary components of delivery time, working time and waiting time, and why waiting time might be the more important one for you to manage.

I'm also gonna talk about three key strategies for reducing your waiting time and therefore accelerating delivery while keeping your quality standards high. And then I'm gonna talk a little bit about how to structure your workflows to keep legal matters working so that clients get their results faster, and then you can improve your practices' cash flow and overall sustainability.

All right, so to start out, let's talk about those two primary components of the overall delivery time, working time and waiting time. And the definitions

are pretty obvious, but I'll call them out anyway. Working time is the actual time you spend actually working on the matter. It's the research, the drafting, the analysis, things like that. Whereas the waiting time is the time where the case or the matter or the deliverables are inactive in your system, waiting for something to happen.

And to illustrate them, I'm going to use a different metaphor. I'm going to introduce a travel metaphor. So booking flights. You probably prefer direct flights because they get you to your destination faster. And if you have to make a connection, you probably want the layover, a.k.a. the waiting time, to be as short as possible, assuming you don't miss a connection. And that's because you're not booking plane tickets for the joy of being in transit. You want to arrive at your destination.

And obviously the same thing is true for legal matters, right? Clients want the resolution, not the extended journey. Now if you're fortunate enough to get one of those direct flights, then a faster airplane means a faster trip, and that makes sense. But if you've got a connection or a couple of connections, then any gains from the faster plane can just get eaten away from time spent sitting around in the connecting airport. Now think about your legal work, right? The things you do for your clients.

I'm gonna guess there aren't a lot of direct flights, meaning there's not times when a client can request an outcome and you can just deliver that work for them in a single sitting from start to finish. If you could, then a faster airplane, meaning speeding up your working time, would be a great investment. But for most legal work, there are a lot of connections before you can get the client to their destination. That's the waiting time.

And I actually want to break down waiting time even a little further into two high-level kinds of waiting. The first is internal waiting, which is delays caused within the firm, within your practice. Usually this is where a deliverable is waiting on a resource, and usually that resource is a person,

to have the available capacity to sit down and do the work needed to get the matter to done, or at least to its next milepost along the way.

There's also external waiting, which are delays outside of your firm's control, right? Waiting on courts, clients, third parties, etc. The important thing to keep in mind though is that clients don't distinguish between these. They just experience the delay. If they're sitting in the airport, they don't know if they're waiting on a plane because it's in transit from some other place, because it's getting worked on for some mechanical issue, or maybe because there's not a pilot or a crew available to fly it. They just know they're stuck in the airport.

So what can we do to reduce waiting time in our law practice? I'm going to give you three high-level strategies, and some of them harken back to earlier episodes. The first one is to limit concurrent work, which I've talked about before as limiting work in progress or limiting WIP. And you can go back just a few weeks to episode 56 for a deeper dive on why WIP limits boost your overall productivity.

But here's the high-level piece. The more balls you're juggling at once, the more you have to divide your time and attention across those balls and put effort into keeping all of them in the air. And that means you have less capacity available to spend on any one ball. To put it another way, and this is the basis of Little's Law that I've talked about before, the more concurrent work you have in progress in your system, the longer it will take to deliver each individual piece of work.

And I've used the freeway analogy before. Once you hit a critical mass and there are too many cars on the freeway at once, every car gets slowed down and nobody is getting home in time for dinner. So when you're limiting WIP, what you're saying is you're going to monotask instead of multitask. To use the Kanban catchphrase, you're going to start less in order to finish more.

Ep #61: Delivering Your Legal Work:
Direct Flight or Long Layover?

As I said before, you don't necessarily need to get the matter all the way to done, but you should focus on delivering all of the things required to get it to its next milepost, something I sometimes refer to as the next natural resting state. In practice, what this usually means is getting it out of an internal waiting state and to an external one, maybe back to the client for review, over to an opposing party or opposing counsel, or filed with a court or agency or whatever tribunal you're in front of.

And I'll admit, this is something that takes some practice, but the best way to reduce WIP is to be intentional about having a WIP limit, right? Actually capping the number of things you're going to do at any one time. And then also you need to put some intentional buffers into your system, some cues upstream of that WIP limit, where a promise deliverable can sit in a managed way until someone has enough available capacity to then sit down and work that thing all the way to done, ideally in a single sitting.

The next thing we want to do to reduce waiting time is to reduce rework and the sort of back and forth iterations that sometimes go into getting a deliverable out the door. And one of the best things you can do to reduce rework is to actually define the deliverable, right? Make it clear to everyone involved in the process that number one, we are creating something tangible that we will actually deliver to someone. We're not just working on something. And then number two, to actually put some definition around what quality looks like for that deliverable.

And I've talked in the past about the Kanban principle of making policies explicit. In this case, we are talking about a very specific type of policy called a quality standard. And really what we're trying to do is improve the first pass quality of the work, right? And this is where that notion of sufficiency comes in that I mentioned a few minutes ago.

One of the other concepts we talk about in Agile is the notion of something being fit for purpose. And that's different from saying something is high

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

quality. The overall goal is to build the deliverable to exactly the specification needed to accomplish the result. We don't want to under build it and have it fall short of the needed quality, but it's just as bad from a process standpoint to overbuild it, to waste some of your finite capacity doing work that isn't necessary to accomplish what your client is trying to accomplish.

This is something a lot of lawyers really struggle with, and I actually think it can be a big contributor to the expense for a lot of legal work. His lawyer's overbuilding their deliverable, but still billing the client on an hourly basis for that work that they're doing. And either way, right, let's say you don't bill the client for that work. Well, then you're just eating the cost, and that's not a good outcome either.

So instead, we want to define some clear quality standards around what that deliverable needs to include. We have a couple of terms for this in Agile. One of them is creating acceptance criteria, and the other is establishing a definition of done, and either one of them can work. Then once we have that quality standard, we want to get really clear about our process for creating a deliverable to that standard, to the acceptance criteria. And this is especially true if multiple people are involved in the process.

We want the drafter to be able to draft the deliverable to spec. And we want the reviewer to be able to evaluate the quality of the work against that spec. And by doing this, we then reduce the handoffs, right? We eliminate some of the back and forth and the rework so that the work is always marching towards that moment of delivery when the value actually becomes useful and noticeable, especially to the client.

Not all handoffs can be eliminated, and handoffs are okay, but you'll drive a lot more efficiency if you make sure the handoffs are moving the work forward and not pinging them back and forth multiple times. So look for

opportunities to reduce sort of the unnecessary revisions and the approval delays. The third thing we wanna do is to proactively manage external waiting. And not all external waiting is the same, right? For some external delays, like waiting on a court or an agency, there's not a lot you can do except manage client expectations as best you can.

Really, your best strategy for those kinds of extended delays where you don't have control is to deliver your own work as quickly as you can, so you get it into their queue sooner.

But there are other types of external weightings that can be managed. I actually talked about the client homework bottleneck and how to get clients to do their part back in episode 55. And that one's worth a listen if you didn't catch it when it came out. But there are actually similar things you can do to better manage opposing counsel and even opposing parties, sometimes your vendors. These are areas that are outside of your direct control, but they're inside your sphere of influence.

I'll give a shout out here to a long ago colleague of mine, Nada Alnajafi, who blogs at ContractNerds.com and actually has a book called Contract Redlining Etiquette that is at least partly about keeping your contract negotiations clean and professional so that both parties are more likely to get what they need out of the end result and achieve that end result faster.

But the bottom line is you wanna make it as easy as possible for your clients, your vendors, your opponents, and any other stakeholder to do their piece of the work in a timely and efficient manner. Not necessarily for the sake of making their job easy, although that's not inherently a bad thing, but because making their job easy is more likely to move your client's matter forward, and that's what we want. It's certainly what your client wants.

I do want to come back for a minute to working time and the ability of technology to accelerate that part of the overall delivery time. And there's

two main areas where I think tech really helps. The first actually goes back to before the generative AI revolution to straight up document automation. And I won't dive deep on this, but I want to be clear that the number one reason to use document automation isn't that it makes the work go faster.

It's that your document templates serve as a quality standard, which makes delivering multiple copies of fit-for-purpose documents more consistent and more predictable, which is another way of saying it makes them less likely to fail quality control or to need expensive rework and other back and forths before getting them out the door.

And one of the cool uses of generative AI here, and from what I understand, Gavel does this and probably some other document automation tools do too, is they can help you strip away the client-specific info from some of your past work product to give you a clean starting point for creating those document templates. And if I understand it correctly, it can also suggest some of the fields and other variables that would go into the necessary parts of your template. It can't necessarily templatize the whole thing for you, but it can give you a really good headstart.

The other reason to speed up working time, including again, reducing the back and forth and the rework time, is that the main reason that any one piece of work is waiting on a resource to come available, that internal waiting, is usually that the resource is doing other work. So there's obviously a role where speeding up the working time can be a tool for reducing wait times. But I want you to be really intentional that that's what you're doing. You're setting up your systems and workflows to get deliverables out the door, not just trying to work on them faster.

All right, so I'm going to leave it at that for this week. But just a quick recap. AI, other technology tools, they're great for reducing the working time. But the real workflow efficiency comes from managing flow across your entire system, not just optimizing certain parts of it. We don't want you to create a

more productive way of doing the work only to have the deliverable itself get stuck at some other non-value-added place, which means from a client experience, they're still waiting on the darn thing, regardless of what the reason for the waiting is.

And really, the takeaway is to be intentional about your overall workflow design around improving the client experience and the case outcomes. The whole system is not just the sum of its parts, it's how those parts interact and work together to deliver the thing that you're trying to do.

If you wanna go a little deeper on how to measure for the flow of work in your practice, and then better manage both the working time and the waiting times that make up the overall delivery time, you can actually sign up for my Agile Attorney Bootcamp email series.

I'm currently in the process of turning this into my next pocket guidebook, but for now you can go to agileattorney.com/resources and look for the Bootcamp sign up link. And in that email series, there are some specific calls to action for you and your team to make your work product delivery smoother and more predictable, which ultimately is going to lead to you and your team delivering the work faster.

All right, that's it for this week. If you found this episode useful, please forward it to a colleague or a co-worker. I also always appreciate it when folks leave me a review in either Apple Podcasts or Spotify, and it does help other people find this podcast.

If you have questions about any of the concepts I've talked about, you can reach out to me at john.grant@agileattorney.com. You can message me on LinkedIn, or actually coming soon, you're gonna be able to ask it inside the free Agile Attorney Community Portal, which you'll find at community.agileattorney.com. More on that in future episodes.

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

Ep #61: Delivering Your Legal Work: Direct Flight or Long Layover?

As always, this podcast gets production support from the fantastic team at Digital Freedom Productions, and our theme music is Hello by Lunara. Thanks for listening, and I'll catch you next week.