You don't get to know what you're fighting for

17 MAY 2015

A number of my recent posts may have given you the impression that I know exactly what I'm fighting for. If someone were to ask you, "hey, what's that Nate guy trying so hard to do," you might answer something like "increase the chance of human survival," or "put an end to unwanted death" or "reduce suffering" or something.

This isn't the case. I mean, I am doing those things, but those are all negative motivations: I am *against* Alzheimer's, I am *against* human extinction, but what am I *for?*

The truth is, I don't quite know. I'm for *something*, that's for damn sure, and I have lots of feelings about the things that I'm fighting for, but I find them rather hard to express.

And in fact, I highly doubt that *anyone* knows quite what they're fighting towards — though it seems that many people think they do, and that is in part why I'm writing this post.

When I wrote on rationality, one commenter replied:

> *I would just note upfront that*

> *Reasoning well has little to do with what you're reasoning towards.*

and

> *Rationality of this kind is not about changing where you're going, it's about changing how far you can go.*

are white lies, as you well know. It's not unusual in the process of reasoning of how to best achieve your goal to find that the goal itself shifts or evaporates.

"How to best serve God" may result in deconversion.

"How to make my relationship with partner a happy one" may result in discovering that they are a narcissistic little shit I should run away from. Or that both of us should find other partners.

"How to help my neighborhood out of poverty" might become "How to make the most money" in order to donate as much as possible.

This is a fine point. Humans are well-known for their ability to start out pursuing one goal, only to find that goal shift drastically beneath them as their knowledge of the world increases. In fact, this is a major plot point in many stories (such as, say, The Foundation Trilogy, The Dresden Files, and The Neverending Story). The goal you think you're pursuing may well not survive a close examination.

I claim this is true even if you think your goals are simple, objective, obvious, high-minded, or sophisticated. Just as the

theist setting out to do the most good might deconvert after deciding that they would still want humanity to flourish even without a divine mandate, so may the utilitarian setting out to do the most good discover that their philosophy is incoherent.

In fact, I suspect this is *inevitable*, at least at humanity's current stage of philosophical development.

It's nice and clean and *easy* to say "I'm a total hedonic utilitarian," and feel like you know exactly what you value. But what does it mean, to be a utilitarian? What counts as a mind? What counts as a preference? Under whose interpretation, under whose process, are preferences extracted? Do you feel an obligation to create people who don't exist? Does a mind matter more if you run two copies of it side by side? I doubt these questions will have objective answers, but subjective resolutions will be complex and will depend on what we value, in which case "total hedonic utility" isn't really an answer. You can *say* you're fighting for maximum utility, but for now, that's still a small label on a complex thing that we don't quite know how to express.

And even if we could express it, I doubt that most humans are in fact total hedonic utilitarians. Imagine that an old friend of yours eats a sandwich which (unexpectedly) alters their preferences so that all they want to do all day is stare at a white wall and not be disturbed. Do you feel a moral obligation to help them find a white wall and prevent others from disturbing them? If there was a button that resets

them to as they were just before they ate the sandwich, would you press it? I sure as hell would — because I feel loyalty not only to the mind in front of me, but to the *person*, the *history*, the *friend*. But again, we have departed the objective utilitarian framework, and entered the domain where I don't quite know what I'm fighting for.

If I am loyal to my old friend over the person who sits in front of the white wall, then am I also obligated to "save" people who naturally want to wirehead? Am I obligated to the values they had as a teenager? Am I obligated to maximize the utilities of babies, before they grow up?

I'm not saying you can't answer these questions. I'm sure that many people have. In fact, I'm sure that some people have picked simple-enough arbitrary definitions and then bitten all the associated bullets. ("Yes, I care about the preferences of rocks a little bit!" "Yes, I maximize the utility of babies!", and so on.) And I'm picking on the utilitarians here, but the same goes for the deontologists, the theists, and everybody else who thinks they know what they're fighting for.

What I'm saying is, even if you *say* you know what you're fighting for, even if you *say* you accept the consequences and bite the bullets, *it's possible for you to be wrong about that.*

There is no *objective* morality writ on a tablet between the galaxies. There are no objective facts about what "actually

matters." But that's because "mattering" isn't a property of the universe. It's a property of a *person*.

There *are* facts about what we care about, but they aren't facts about the stars. They are facts about *us*.

There is no objective morality, but also your morality is *not* just whatever you say it is. It is possible for a person to say they believe it is fine to kill people, and *be lying*. The mind is only part of the brain, and it is possible to have both (a) no objective morality, and (b) people who are wrong about what they care about.

There are facts about what you care about, but you don't get to know them all. Not by default. Not yet. Humans don't have that sort of introspective capabilities yet. They don't have that sort of philosophical sophistication yet. But they *do* have a massive and well-documented incentive to convince themselves that they care about simple things — which is why it's a bit suspicious when people go around claiming they know their true preferences.

From here, it looks very unlikely to me that anyone has the ability to pin down exactly what they really care about. Why? Because of where human values came from. Remember that one time that Time tried to build a mind that wanted to eat healthy, and accidentally built a mind that enjoys salt and fat? I jest, of course, and it's dangerous to anthropomorphize natural selection, but the point stands: our values

come from a complex and intricate process tied closely to innumerable coincidences of history.

Now, I'm quite *glad* that Time failed to build a fitness maximizer. My values were built by dumb processes smashing time and a savannah into a bunch of monkey generations, and I don't entirely approve of all of the result, but the result is also where my approver comes from. My appreciation of beauty, my sense of wonder, and my capacity to love, all came from this process.

I'm not saying my values are dumb; I'm saying you shouldn't expect them to be simple.

We're a thousand shards of desire forged of coincidence and circumstance and death and time. It would be *really surprising* if there were some short, simple description of our values. Which is why I'm always a bit suspicious of someone who claims to know exactly what they're fighting for. They've either convinced themselves of a falsehood, or they're selling something.

Don't get me wrong, our values are not *inscruitable.* They are not *inherently unknowable.* If we survive long enough, it seems likely that we'll eventually map them out.

But we don't know them *yet.*

That doesn't mean we're lost in the dark, either. We have a hell of a lot of *evidence* about our values. I tend to prefer

pleasure to pain and joy to sadness, most of the time. I just don't have an exact description of what I'm working towards.

And I don't *need* one, to figure out what to do next. Not yet, anyway. I can't tell you exactly where I'm going, but I can sure see which direction the arrow points.

It's easier, in a way, to talk about the negative motivations — ending disease, decreasing existential risk, that sort of thing — because those are the things that I'm pretty sure of, in light of uncertainty about what really matters to me. I don't know exactly what I want, but I'm pretty sure I want there to be humans (or post-humans) around to see it.

But don't confuse what I'm *doing* with what I'm *fighting for*. The latter is much harder to describe, and I have no delusions of understanding.

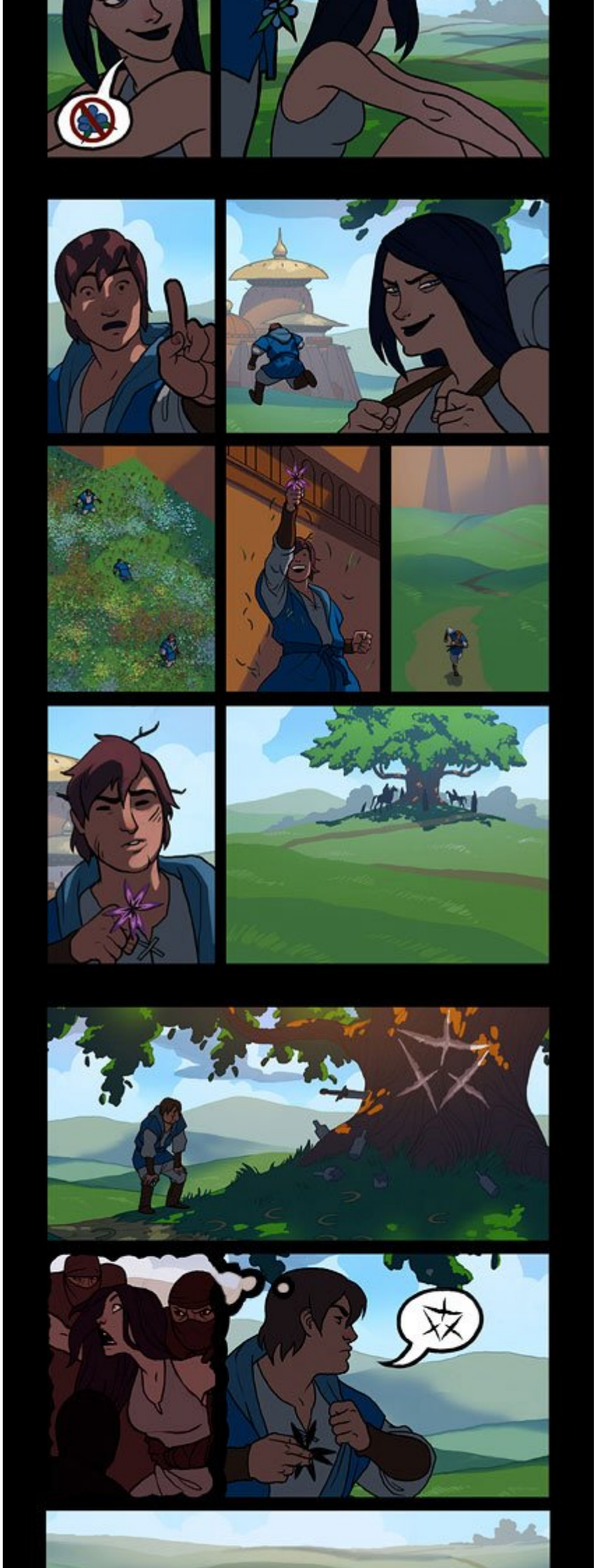
You don't get to know exactly what you're fighting for, but the world's in bad enough shape that you don't *need* to.

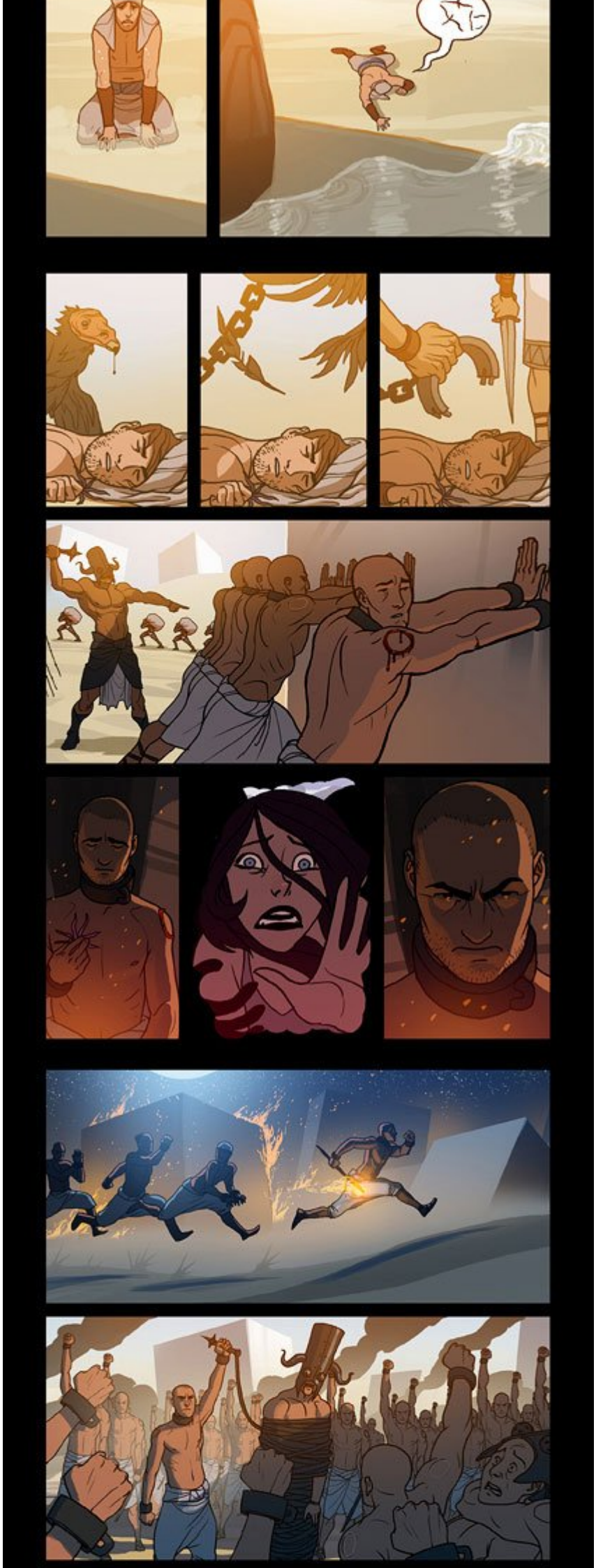
In order to overcome the listless guilt, I strongly recommend remembering that you have something to fight for, but I also caution you against believing you know exactly what that thing is. You probably don't, and as you learn more about the world, I expect your goals to shift.

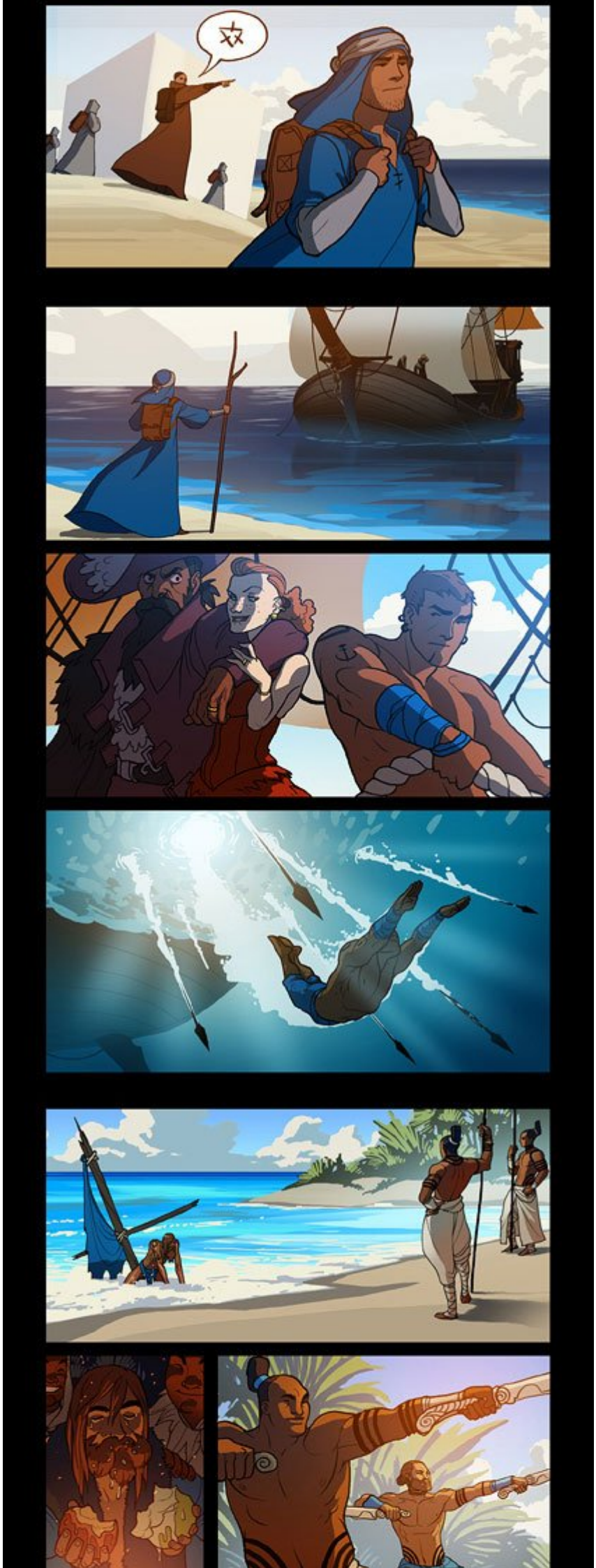
I'll conclude with a comic by Matt Rhodes:

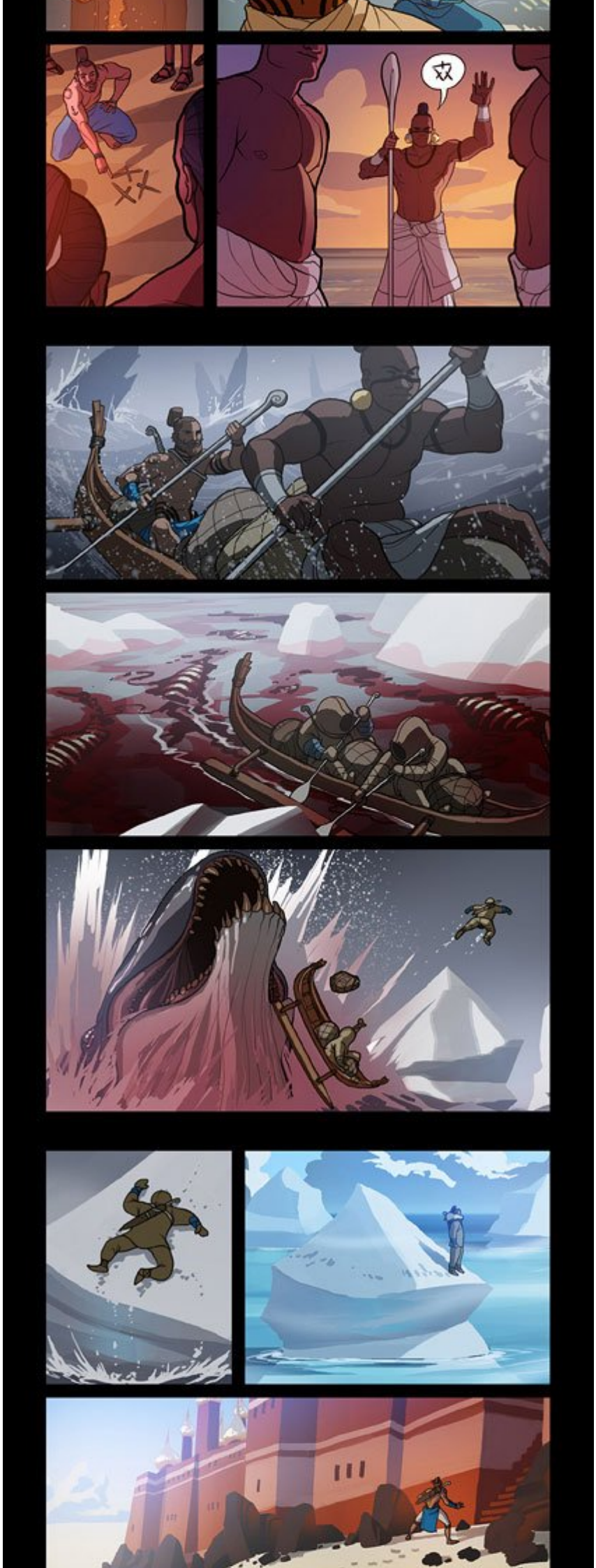
My Hero

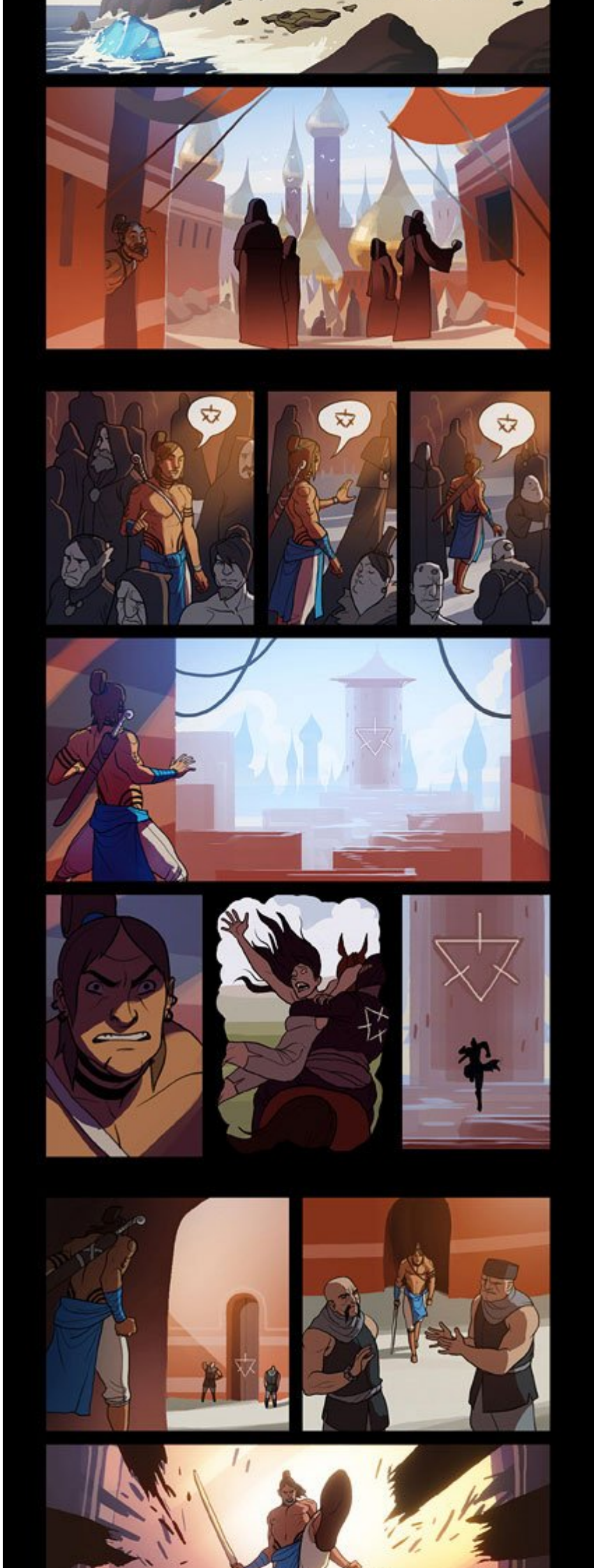
by Matt Rhodes

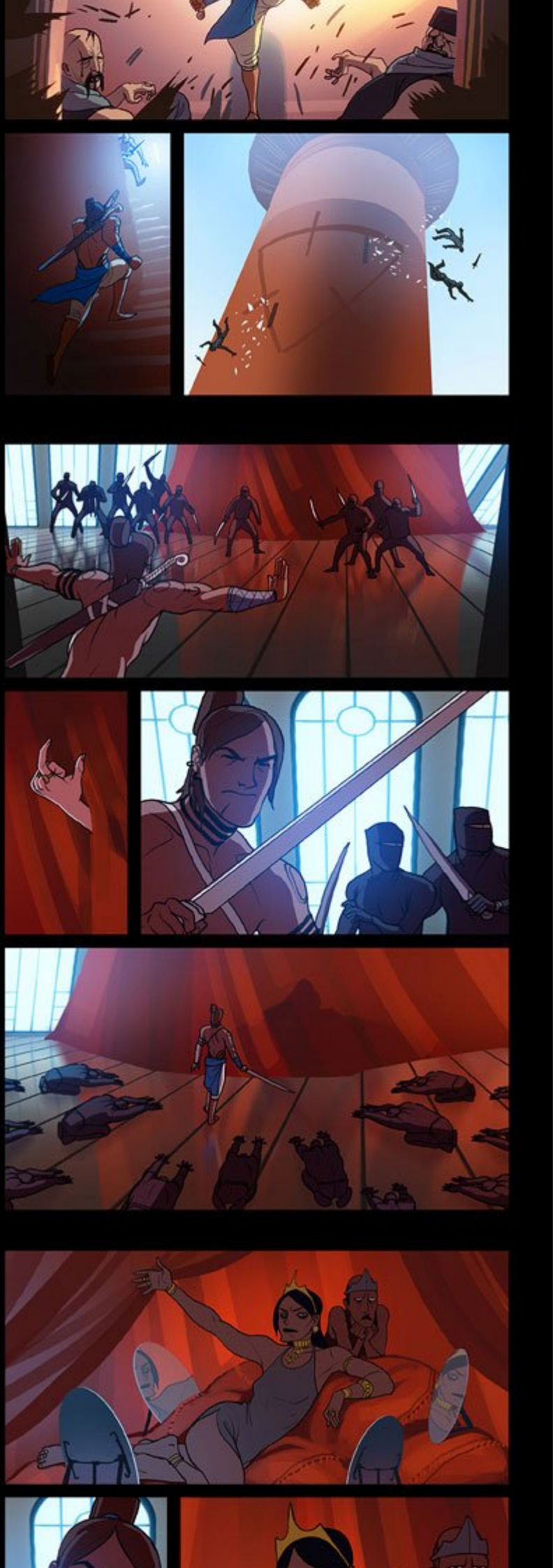


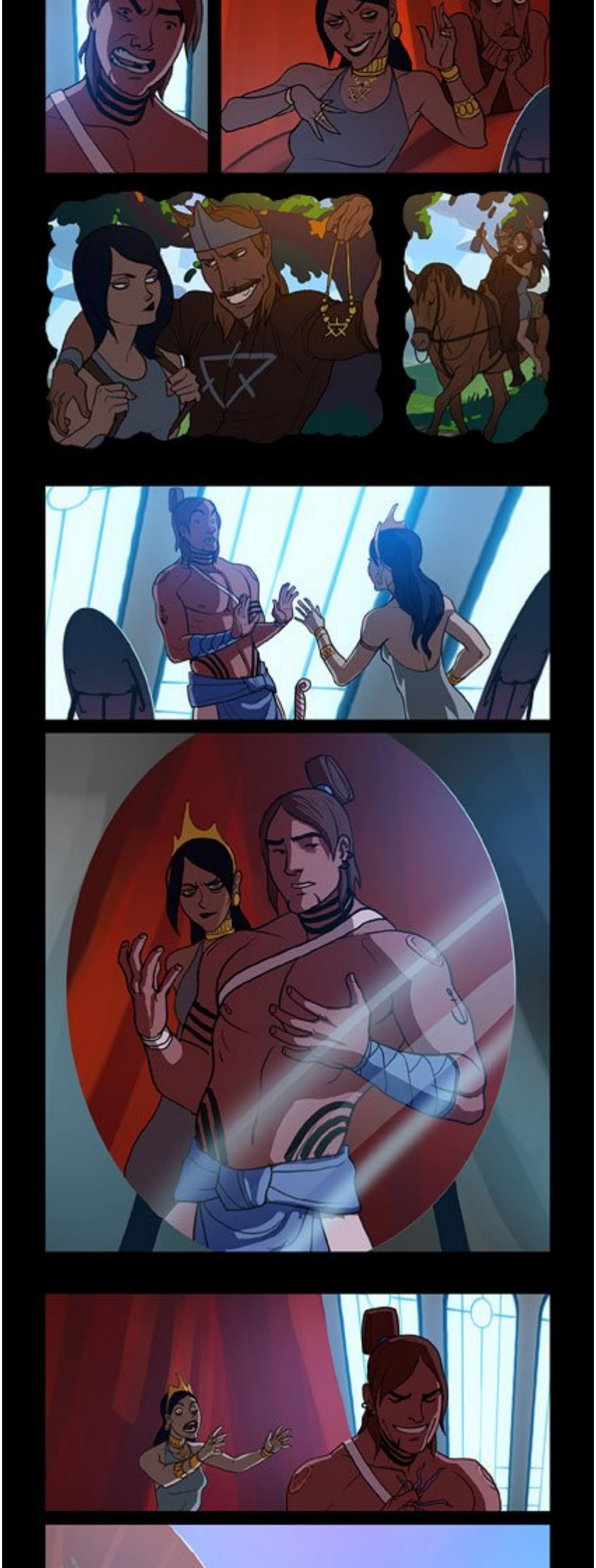


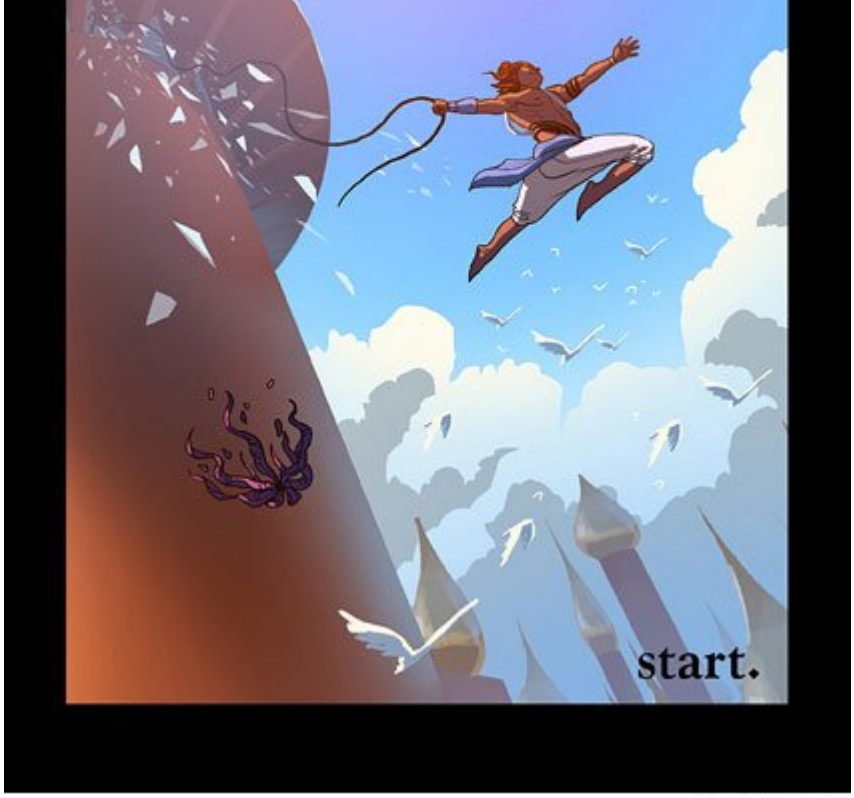












Matt Rhodes

(source)