An Economic Theory of Intermediary Liability

From the DMCA to the DSA:
A Transatlantic Dialogue on Online Platform Liability and Copyright Law
April 7, 2023

James Grimmelmann
Pengfei Zhang

Motivation

- Economic claims about intermediary liability are common:
  - E.g., platform liability creates chilling effects
  - E.g., platforms do/don’t have an incentive to self-police
  - E.g., Section 230 does/doesn’t balance freedom and safety
- But these claims are mostly informal
  - They are policy arguments, not testable propositions

Why an economic model? (for legal scholars)

- Provide a common framework to compare arguments
- Build intuition for important effects and tradeoffs
- Visualize consequences of liability rules
- Make implicit assumptions explicit

Why an economic model? (for economists)

- Prove theorems about efficiency conditions
- Know what econometric questions to ask

In this talk

- Model overview
- What do platforms do if they have ***blanket immunity***?
- What do platforms do if they face ***strict liability***?
- Policy responses to ***undermoderation***: actual knowledge, liability on notice, negligence, and conditional immunity
- DMCA § 512, DSA, CDA § 230

Not in this talk

- Fancy math
- Platform investigations
- Policy responses to ***overmoderation***: subsidies, must-carry

Overview

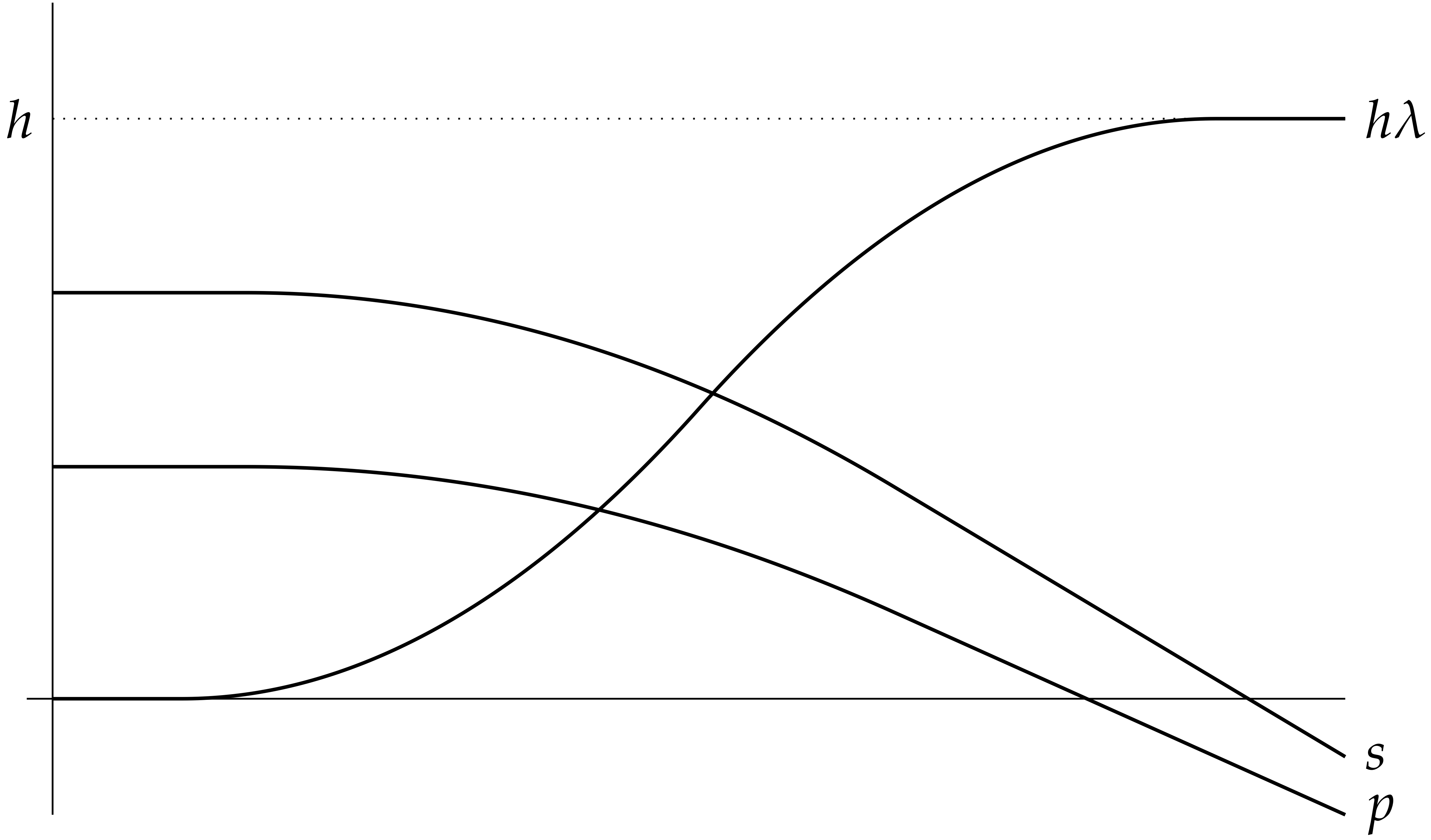
A model of moderation

- Users submit discrete items of ***content*** to a platform
  - Each item is either ***harmful*** or harmless
- The platform choose whether to ***host*** or ***remove*** each item. If it hosts:
  - The platform receives some ***revenue*** p
  - Society receives some ***benefits*** s
  - If harmful, third-party victims suffer ***harm*** h
- The platform ***does not know*** with certainty which items are harmful
  - It observes the ***probability*** λ that an item is harmful

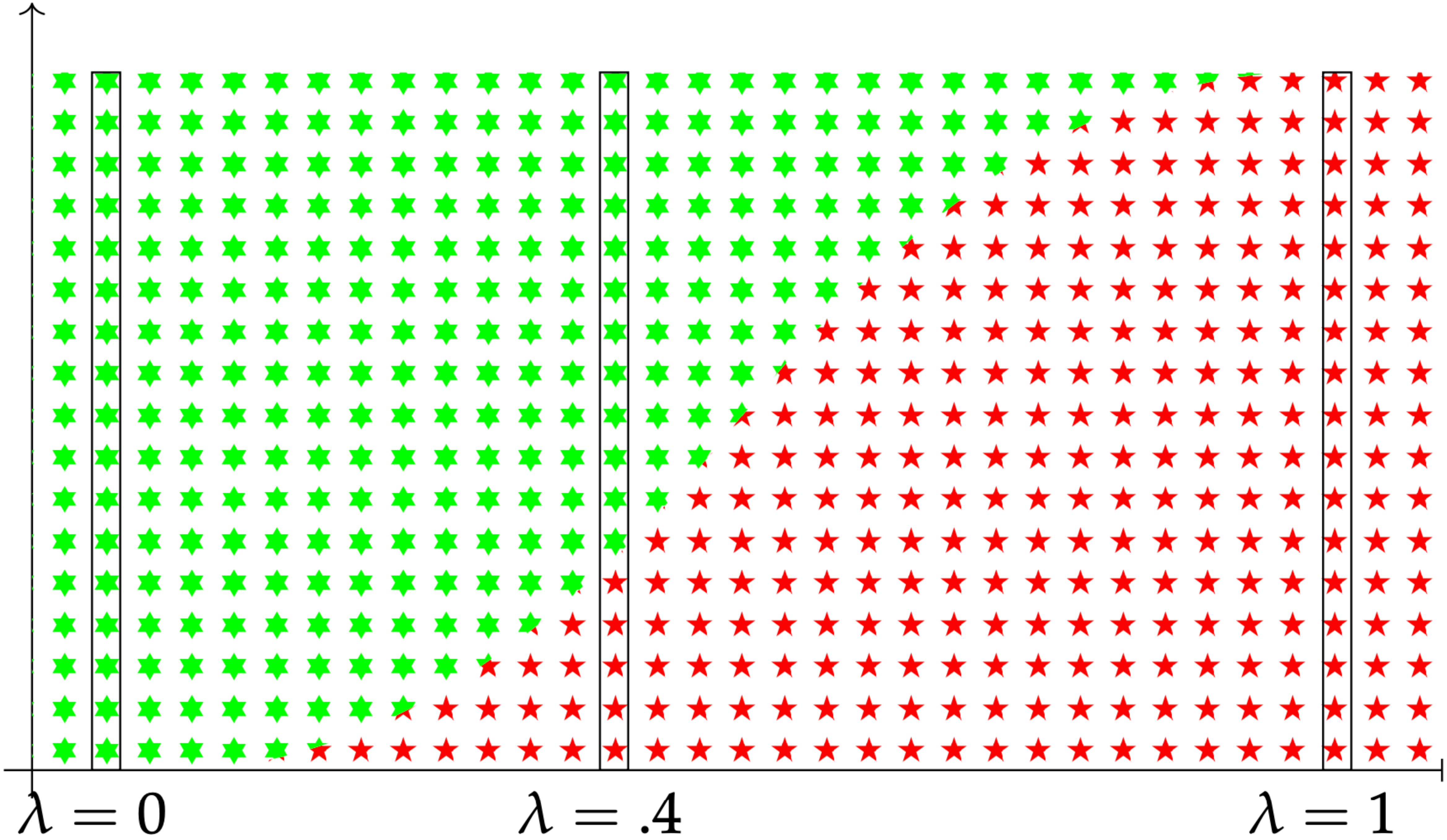
Core assumption

- In reality, p, s, h, and λ are ***complicated functions***
- We simplify them by collapsing content onto a ***single axis***
- As you go from left to right, you go from “good” to “bad”:
  - Content is ***less profitable*** to the platform: p decreases
  - Content is ***less beneficial*** to society: s decreases
  - The harm (if it happens) is ***fixed***: h is constant
  - Content is ***more likely to be harmful***: λ increases

The spectrum of content



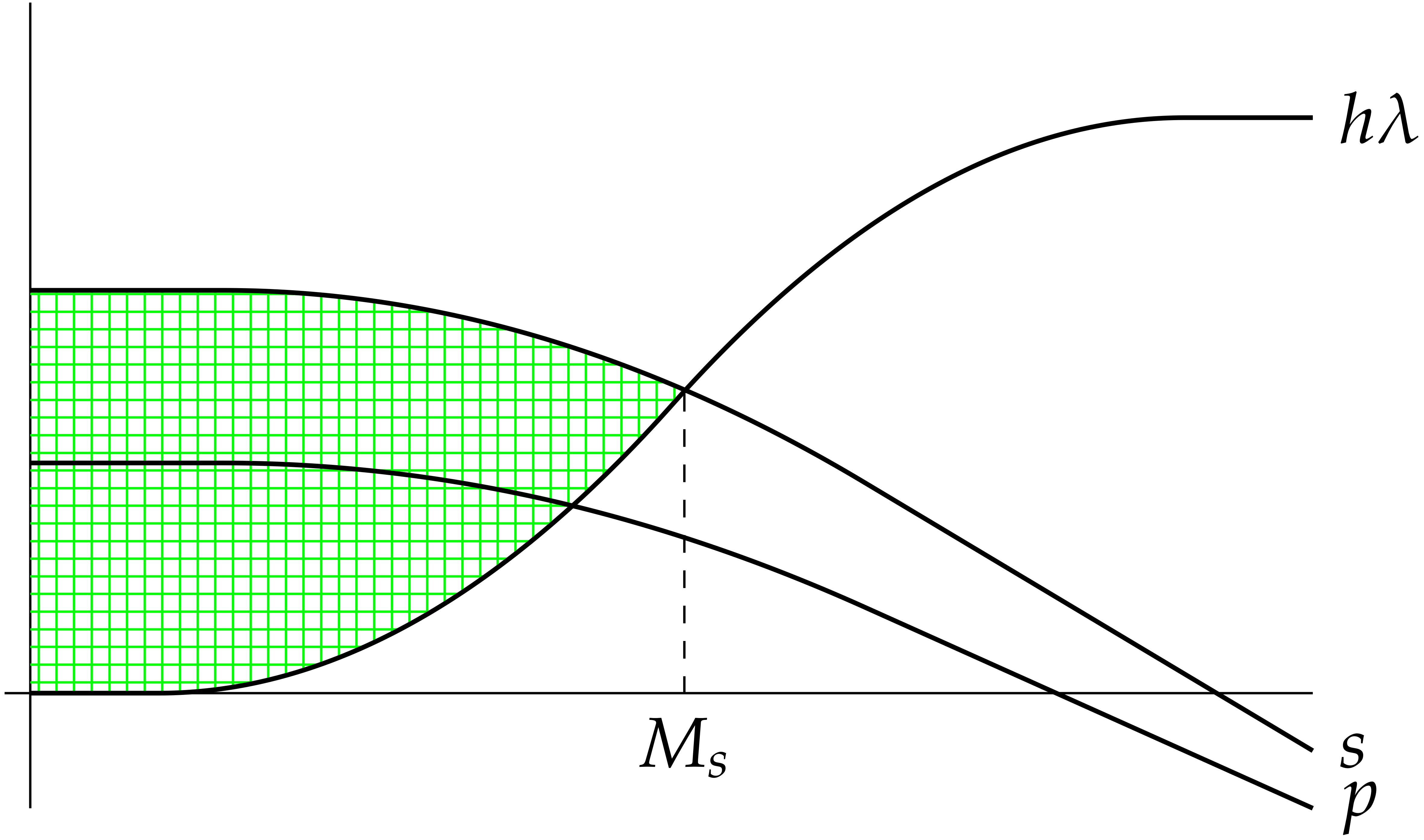
What should a rational moderator do?



Rational content moderation

- Content further to the right is always worse *ex ante*
    - It has lower (known) benefits but higher (expected) harms
- A rational moderator sets a ***moderation threshold*** *M*
    - Content to the left of *M* stays online
    - Content to the right of *M* is taken down
- *M* incorporates the moderator's judgments about the ***acceptable risk of harm***

The ***efficient moderation threshold*** M_s is where the marginal benefits s equal the marginal expected harms $h\lambda$

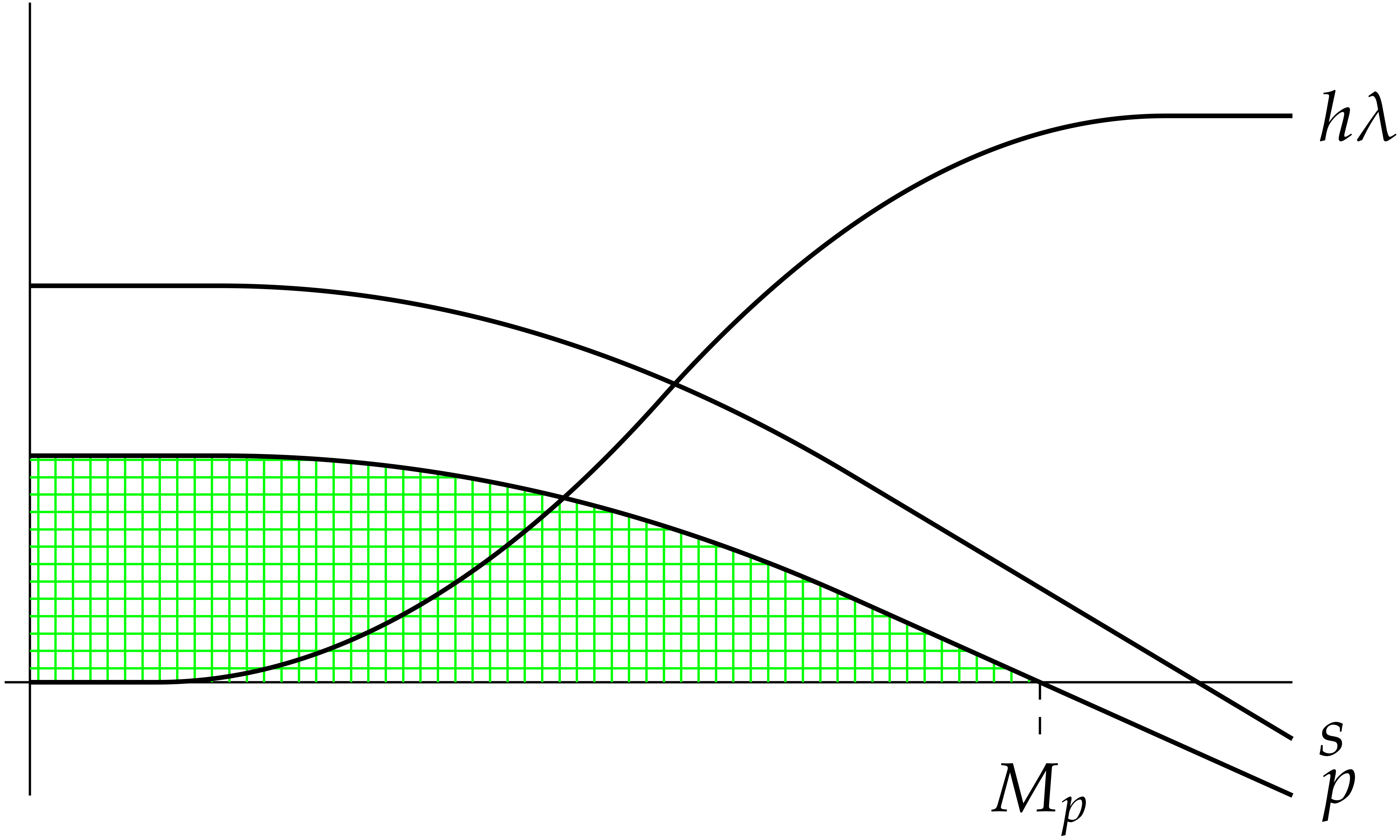


The optimal level of harmful content is not zero

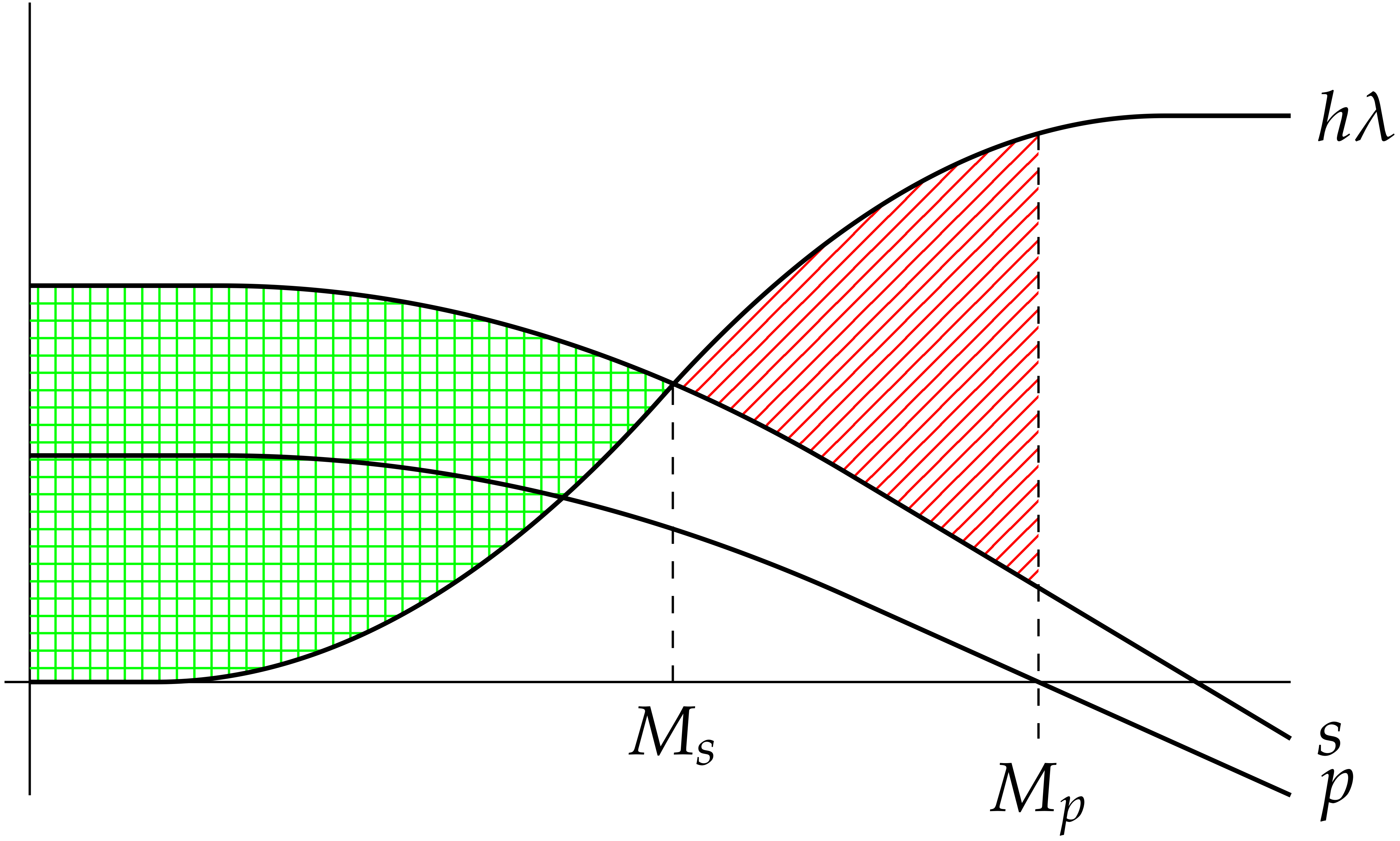
- Any choice of *M* trades off false positives and false negatives
  - High threshold = more “bad” content stays online
  - Low threshold = more “good” content taken down
- We tolerate some harmful content because it is indistinguishable *ex ante* from beneficial content
  - *Users* and *victims* may know whether content is harmful
  - *Platforms* and *regulators* typically have less information

Blanket immunity

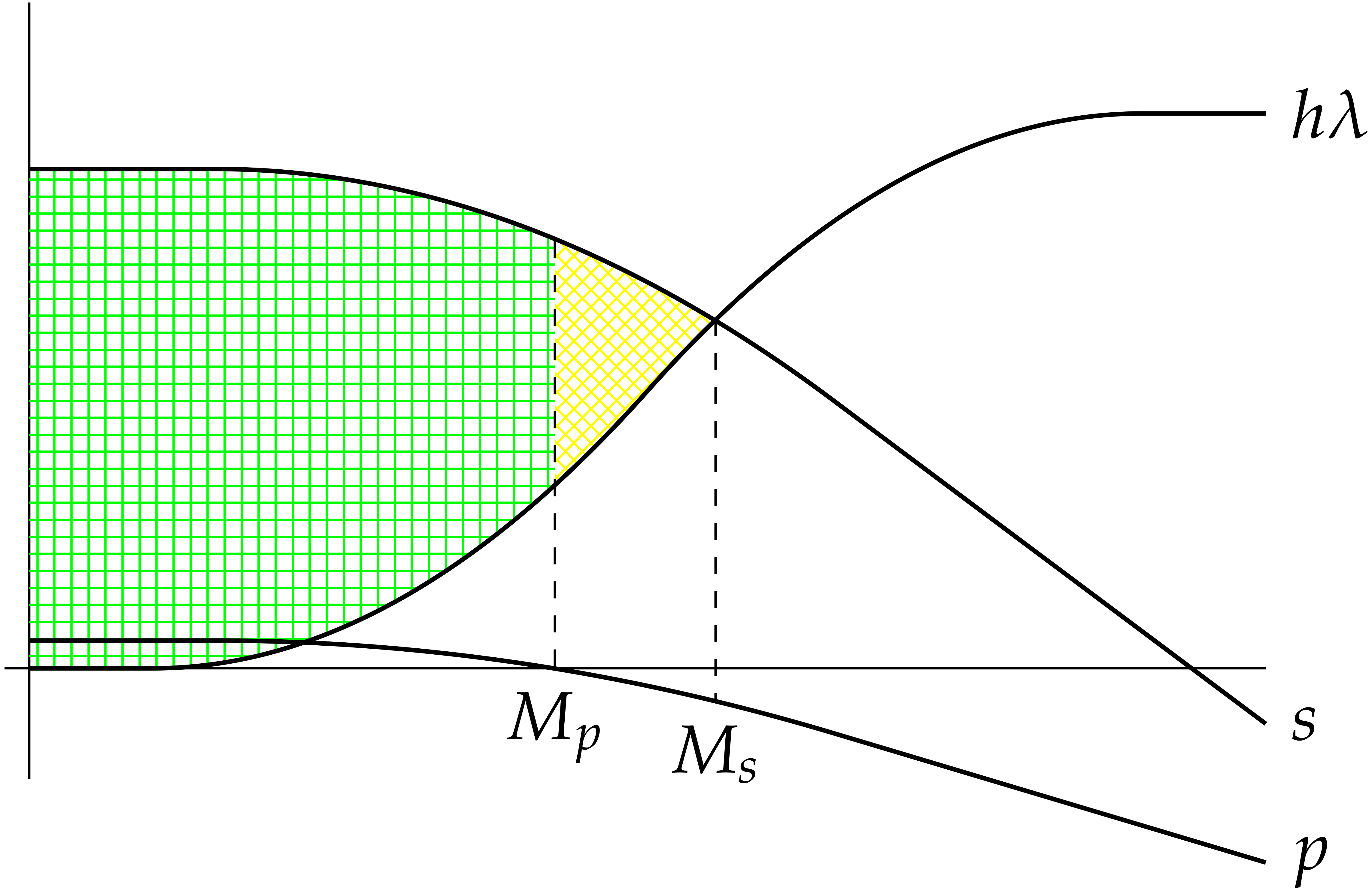
The platform's ***profit-maximizing moderation threshold*** M_p is where its marginal revenue $p = 0$



If $M_p > M_s$ the platform ***undermoderates***

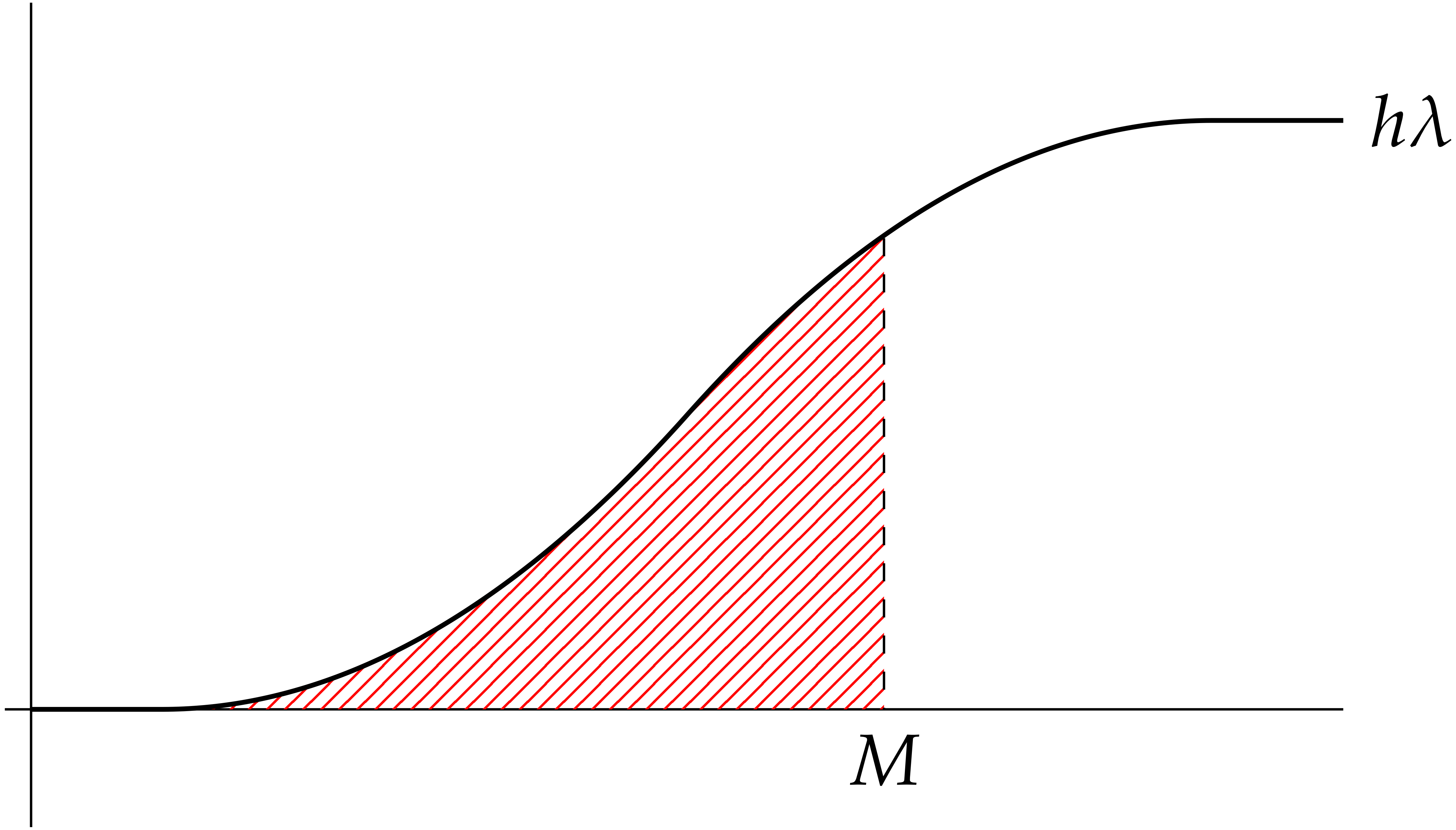


If $M_p < M_s$ the platform ***overmoderates***

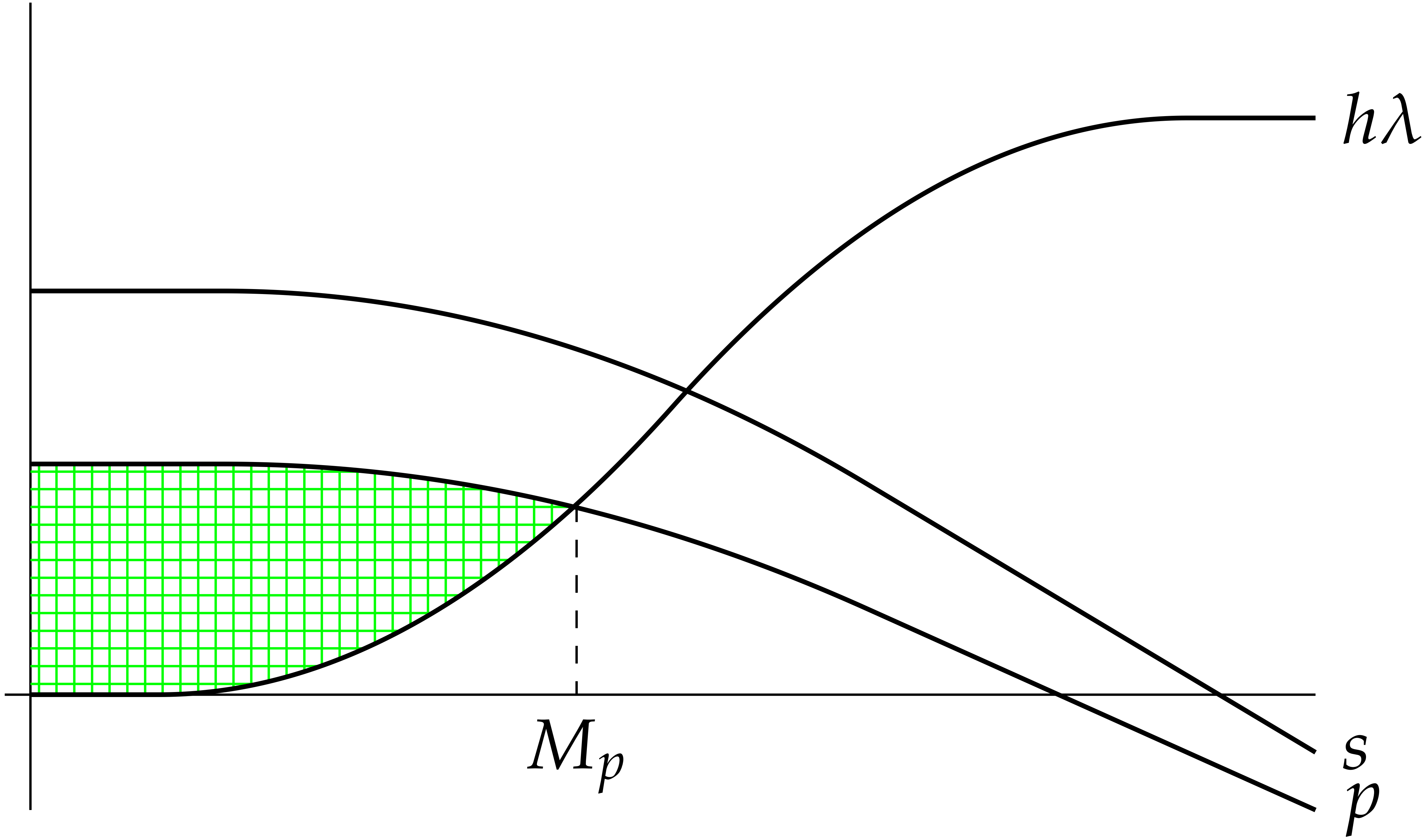


Strict liability

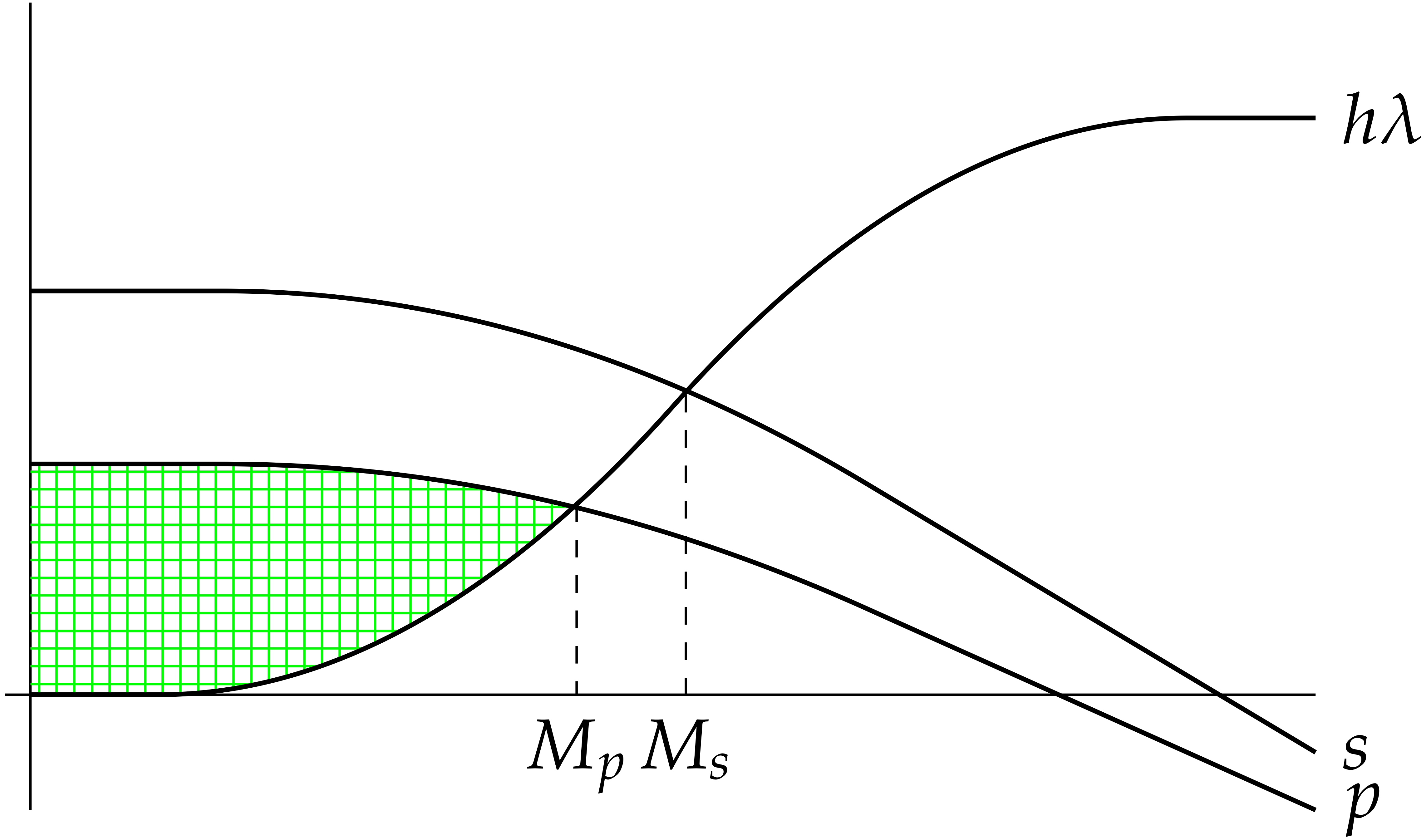
Under ***strict liability***, the platform must pay damages for all harm caused by content it hosts



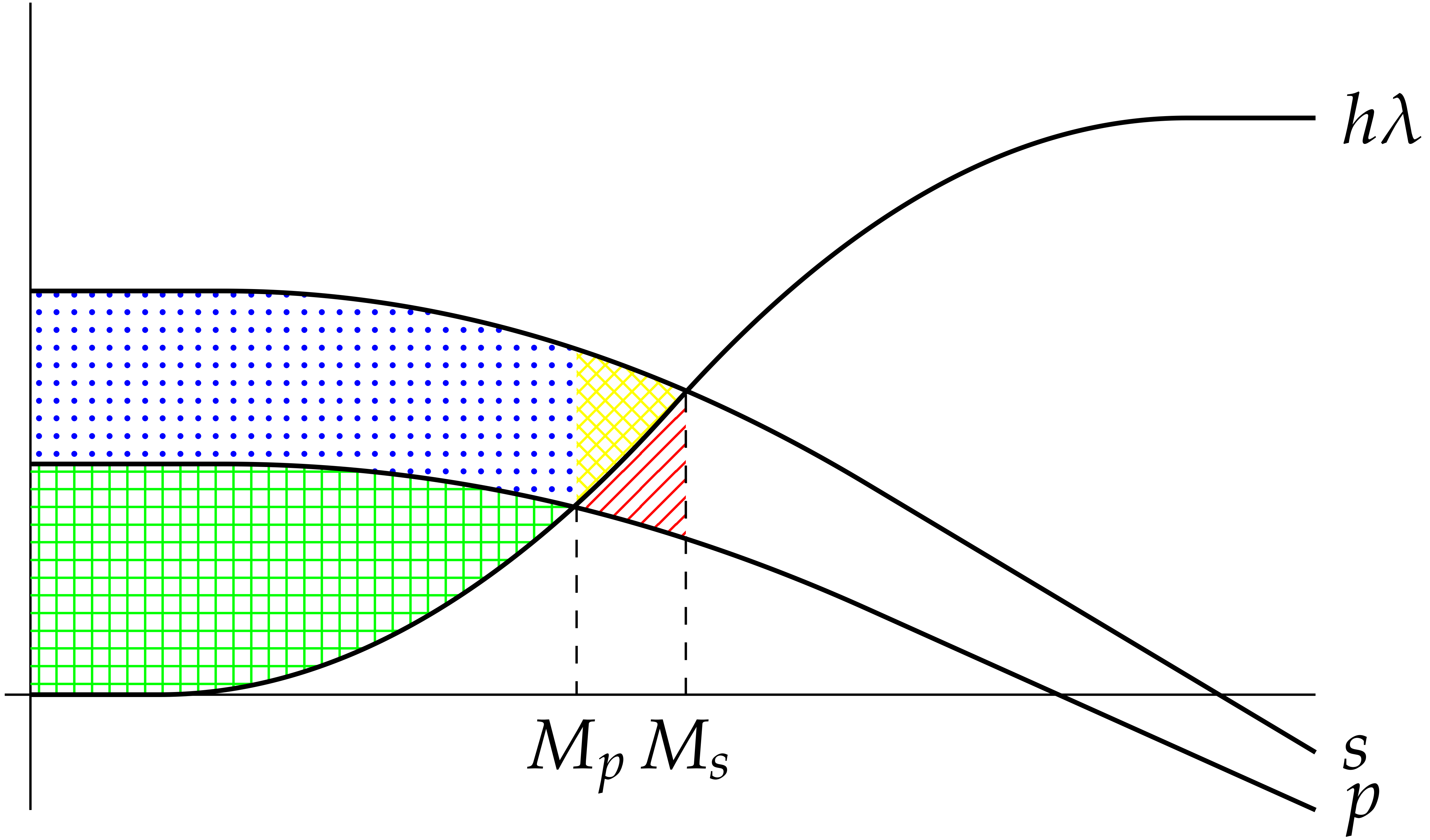
Under strict liability, the platform's ***profit-maximizing moderation threshold*** is when $p = H\lambda$



The platform ***always overmoderates*** under strict liability



Strict liability causes a ***welfare loss***: some content is unprofitable (to the platform) but beneficial (to society)



Collateral censorship

- Felix Wu's theory of collateral censorship has two parts:
  - (1) "good" content has positive externalities
  - (2) "good" and "bad" content are indistinguishable *ex ante*
- If either assumption fails, strict liability is efficient
- But both together can justify intermediary immunity
- Strict liability makes the platform internalize the harms from the content it carries, but not the benefits

Other liability regimes

Actual knowledge

- The platform is liable for an item of harmful content when it ***knows*** that the item is harmful and fails to remove it
    - E.g., DMCA § 512512(c)(1)(A)(i)
- Economic intuition: ***no investigation*** is required
- Implementation note: does “actual knowledge” actually mean actual knowledge?

Liability on notice

- The platform is liable for an item of harmful content when it ***receives a notice*** about the content and fails to remove it
    - E.g., DMCA § 512(c)
    - E.g., DSA art. 16
- Economic intuition: notices lower the cost of ***investigation***
    - Someone else can investigate ***more cheaply***
    - Someone else has a ***stronger incentive*** to investigate

Notice as a signaling game

- Notices work because they ***convey information***
  - Receiving a notice is different than not receiving one
  - "I have investigated this content and it is is harmful."
- But this signal ***need not be true***
  - When investigations are costly, victims will shirk
  - They will send notices ***without investigating***
- Game theory: liability on notice collapses into strict liability

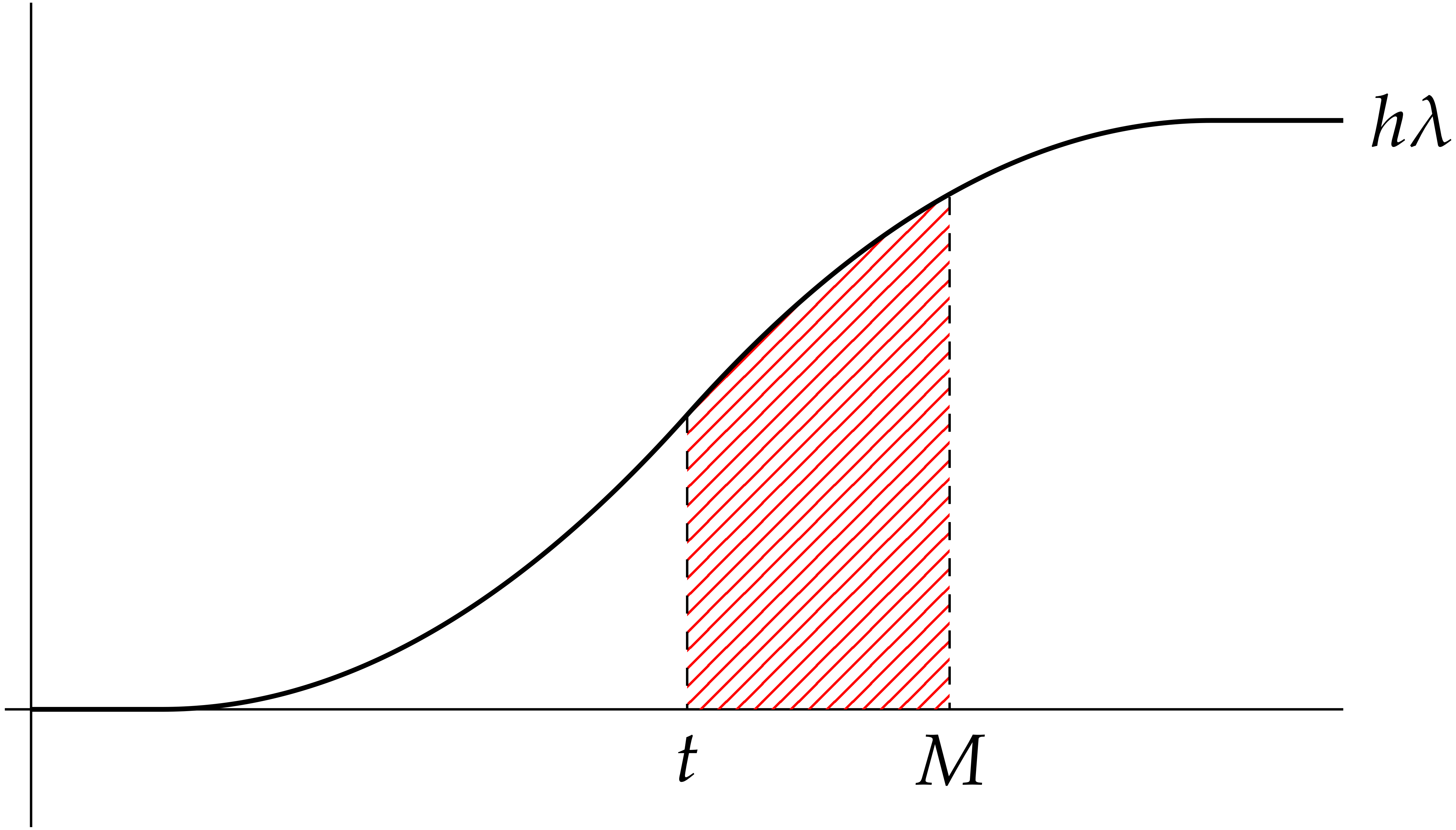
Making liability on notice work

- Key policy response: ***deter sending false signals***
- E.g., ***penalties for sending false notices***
  - DMCA § 512(f), *but see Rossi* and *Lenz*
  - DSA art. 23(2) repeat-grumbler suspensions
- E.g., notices from parties with ***less incentive to shirk***
  - DSA art. 22 trusted-flagger system?

Negligence

- The regulator sets a ***threshold*** of ***probability of harmfulness***
- The platform is liable for content that
  - Was *ex ante* more likely to be harmful than the threshold
  - And *ex post* actually turned out to be harmful
- E.g., DMCA § 512(c)(1)(A)(ii) "red flag" knowledge
- Economic intuition: use liability to promote moderation, while also letting the platform not bother beneath the threshold

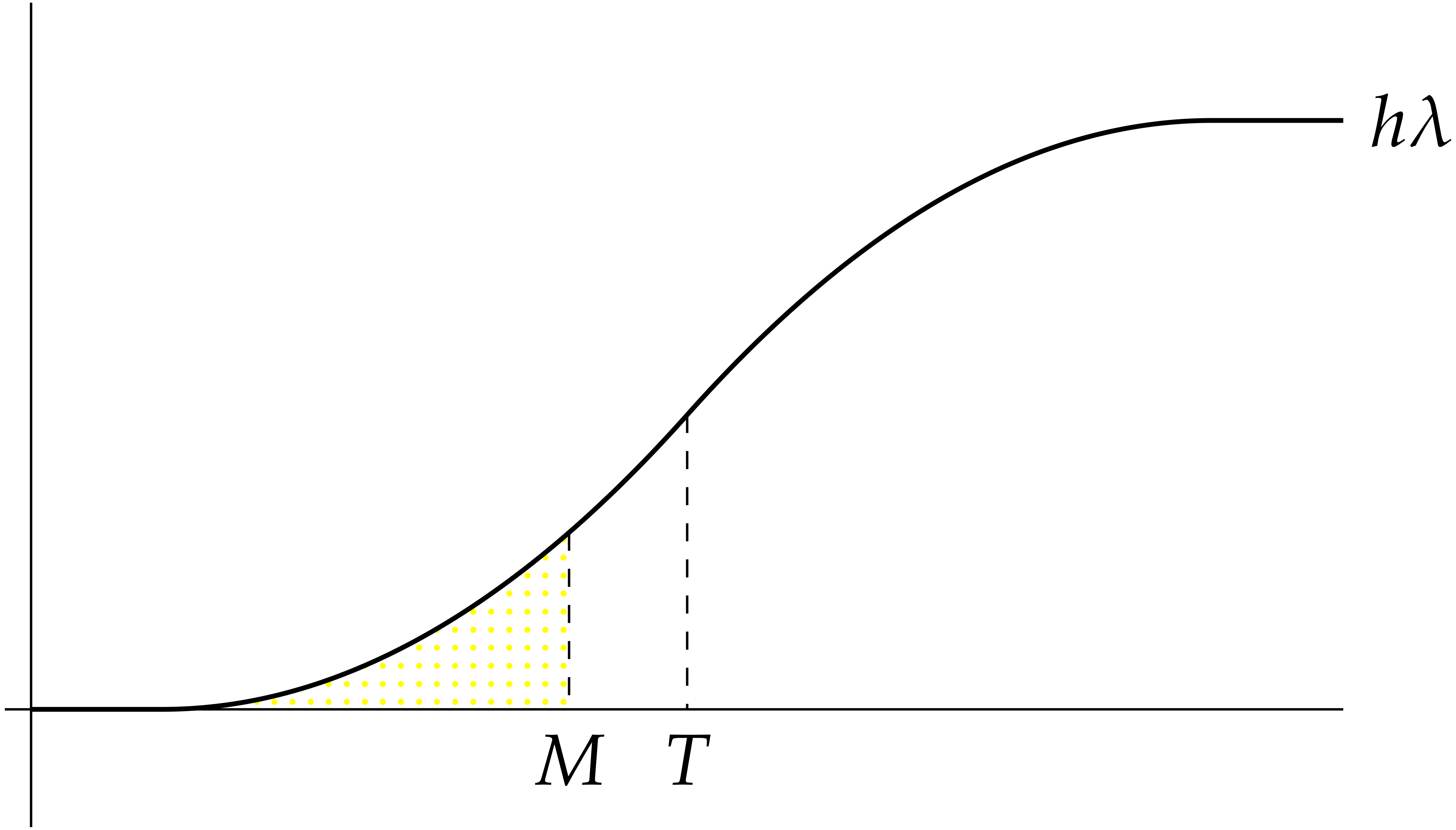
Negligence



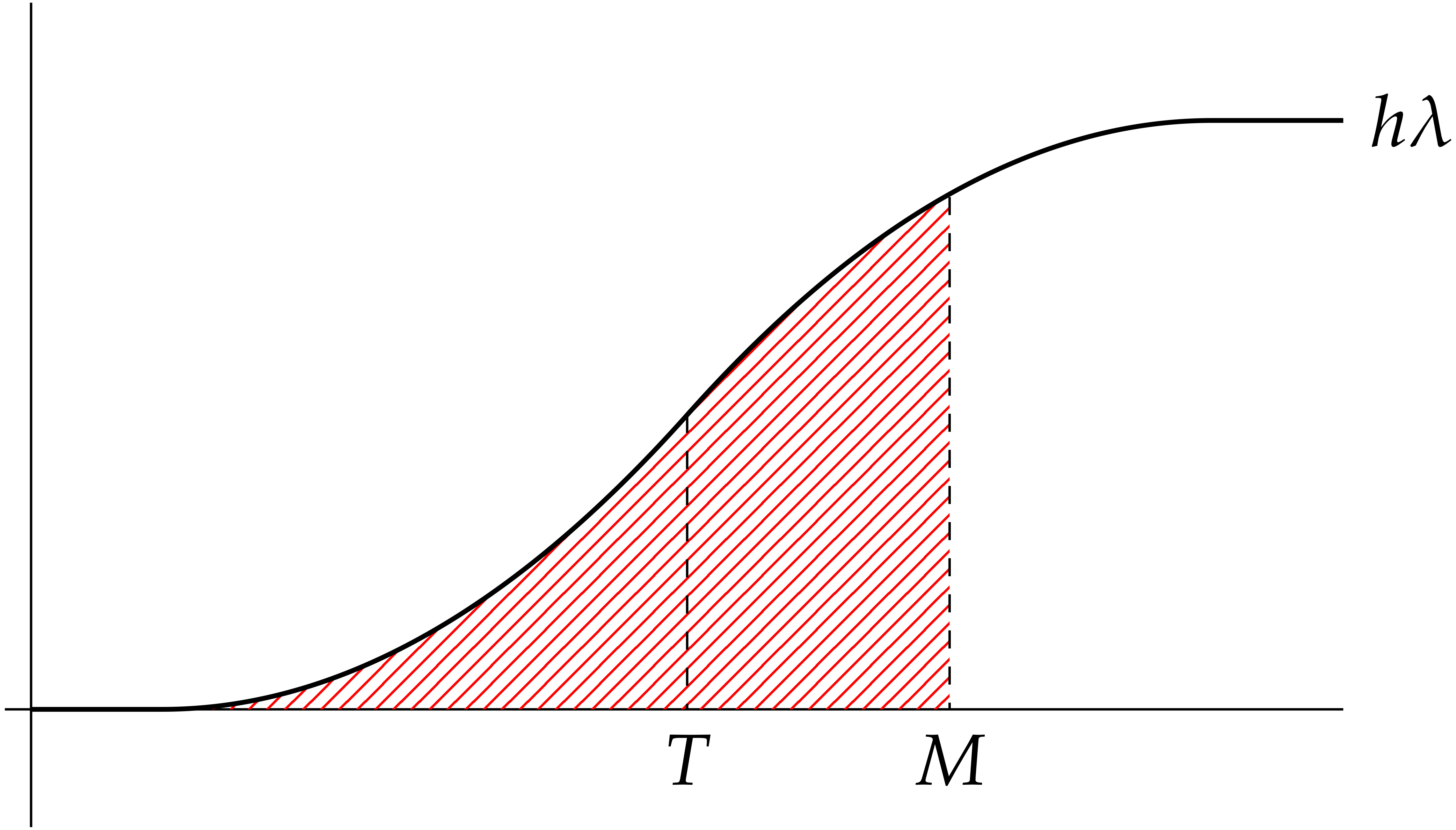
Conditional immunity

- The regulator sets a ***threshold*** of ***total harm***
  - If total harm is below the threshold, the platform is immune
  - If total harm is above the threshold, the platform is strictly liable — even for harms below the threshold
    - E.g., DMCA § 512(i)(1)(B) repeat infringer condition
    - E.g., Citron-Wittes § 230 reform proposal
- Economic intuition: same as negligence!

Conditional immunity (below threshold)



Conditional immunity (above threshold)



Negligence vs. conditional immunity

- Both depend on correct threshold-setting
  - But conditional immunity requires a more comprehensive calculation of harms and benefits over a wider range
- Conditional immunity is discontinuous at the threshold
  - Platforms face severe consequences for getting it wrong
  - See, e.g., *BMG v. Cox*
  - Requires higher confidence in courts' accuracy

Legal regimes

DMCA section 512

- Baseline of immunity, but …
    - § 512(c)(1)(A)(i): actual knowledge
    - § 512(c)(1)(A)(ii) (“red flag”): negligence
    - § 512(c)(1)(B) (“financial benefit”): high p for high λ
    - § 512(c)(1)(C): notice and takedown
    - § 512(i): conditional immunity

DSA

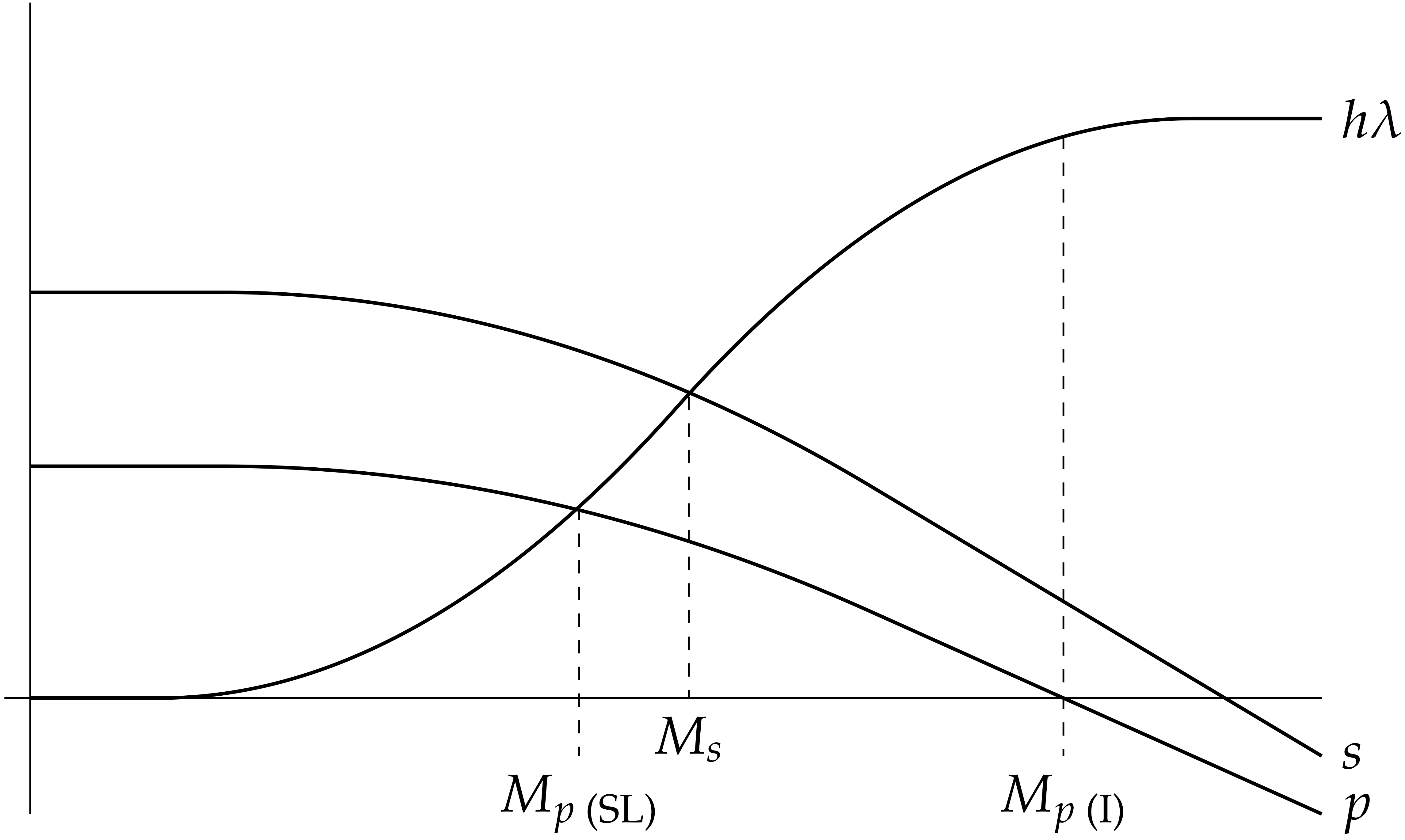
- Baseline of immunity, but …
  - art. 6: actual knowledge *and* negligence
  - art. 9: liability on notice
    - art. 22: trusted flaggers respond to signaling problem
  - art. 23(1): must suspend users “that frequently provide manifestly illegal content”
    - Freestanding obligation, not a conditional immunity

CDA section 230

- Immunity, immunity, immunity, immunity
- Every legal reform imaginable has been proposed:
  - Actual knowledge
  - Negligence
  - Conditional liability
  - Liability on notice

Conclusion

If you only remember one thing from this talk, make it this diagram



A little intuition goes a long way

- Our model is deliberately (and painfully) simplistic …
  - … but it makes the effects of liability rules obvious
- Content moderation is all about threshold-setting …
  - … and so is intermediary liability law

Thank you