Training
This is from an external perspective. I do not personally train these models for ethical reasons, among other things.
6. Visual media, audio content, and written works. These are most often collected by scrapers, automated programs that search all public websites and scavenge any possible content meeting their criteria, regardless of the owner's consent. Far faster than a human could, too.

6.1. Scrapers, mostly, by my understanding. Dunno about whether datasets are collected by third parties first. Scrapers aren't too hard to put together, so I can't imagine they are.

6.2. Basically never, except possibly through the terms of service of a platform they are hosted on. (see Twitter/X) I have never seen a license that explicitly mentions AI use.

6.3. Public domain works that can be easily included in the data set are almost always included.

6.4. Dunno. Possibly, if they wish to be able to re-create or re-train the models later. That's the kind of thing that isn't feasible to tell from an external perspective.

7. You label and feed massive amounts of information into a machine learning implementation in a way that the machine can interpret, try to get it to generate content approximately close to its data set (for example, a picture of an apple when 'apple' is requested) and then reinforce the correct output with a simulated reward and disincentivize incorrect outputs with simulated punishment.

7.1-7.2: I don't know the technical details, but you can find no shortage of information about the process online.

7.3. No. The only way to remove influence from a specific piece of the training set from a model is to completely throw out the model and replace it with a new model that did not include the piece in its training set. Influences can be suppressed and/or hidden with enough punishment, but this is not economically feasible.

7.4. Yes and no. You can ask, for example, ChatGPT to generate a fake Bible quote for fun, and it will output something in the correct format for a Bible quote. Obviously, it was trained on the Bible. But for more specific works with less broad availability, tricking the AI into showing its training is functionally impossible.

8. No circumstances. Unless the model inherits use restrictions from the training data, which runs into enforceability issues, any kind of use-based exceptions or permissions should not apply to AI models. Fair use is a protection of human creativity and innovation, not automated production (or reproduction) of content.

8.1. At a glance, I would not interpret Google v. Oracle America to be relevant. Andy Warhol Foundation v. Goldsmith seems to be more relevant at a passing glance. The purpose and character of the use of copyrighted works should be evaluated by the present use of the model by the model's owners and anyone they knowingly permit to access the model (and in the case of a public service, that they could reasonably have detected and restricted access from)

8.2. The analysis should be done based upon the conditions of distribution and the licensing status of the works with regards to the distributor.

8.3. Present use, as per my answer to 8.1, unless past uses violate fair use.

8.4. Incomprehensible amounts of data when it comes to getting a decent result. No, the volume should not affect fair use status.

8.5. Both the body of works of the same author and the market for the general class of works, especially tending toward the former in cases where style mimicry or significant similarity.

9. Consent should be opt-in. Existing platform terms of use (Twitter/X, for example) already see AI training as distinct from other uses for copyrighted content that is worthy of explicit disclosure. As a result, licensing should have to EXPLICITLY GRANT these permissions for training to be allowed.

9.1. Yes, given how easy it is to pivot a model made for research/education into a commercial product.

9.2. Websites already have things like robots.txt files and APIs which politely ask automated processes (crawlers, scrapers, other services) to behave in certain ways, whether that be to not include a given page in search results or to provide specific information alongside their requests. It would likely need to either be a file metadata thing (controllable by the file's creator) or a similar implementation to the above to function.

9.3. Practically, robots.txt is not enforceable if a scraper wants to ignore it. APIs can return an error if you don't make a request right, but once the request is responded to, there is nothing the server can do. Consent would probably most practically be achievable through an opt-in system where content is explicitly uploaded by creators for the purpose of training.

9.4. Punitive damages and royalties for the lifespan of the model and any models derived from (not trained from scratch) it, destruction of the infringing model in extreme cases or in cases where the appropriate level of royalties given the dataset's size cannot be determined.

9.5. Yes, as you could conceivably entrap a creator by purchasing a notable volume of products from them that you could use to adjust a wider dataset. (see Style Mimicry mentioned by the Glaze project linked in the main comment) It should be covered as part of the rights-transferring agreement. The buyer should be able to object, however, if the creator chooses to retain these rights to their work and use them for that purpose as well.

10. Explicitly. The owner must agree, either by providing a license to their work generally (not unlike Creative Commons) or by agreeing to a licensing agreement provided to them by a third party.

10.1. Should be, if you use a standard template license. Don't see why it couldn't be. Would slow down data collection and such, but not make it impossible.

10.2. Probably. The organizations would need to adapt, but I don't see why they couldn't create departments dedicated to finding infringements by public models.

10.3. Yes. The license should cover the training, use, and distribution of the model, with the royalty rate being set and allocated by the volume of content (in number of works) included in the model's data set compared to the entirety of the copyrighted volume. Distribution method should be handled during the enforcement of the license.

10.4. I'm unsure, tending towards no. If anything, ECL groups should need rights transferred akin to 9.5 to be able to enter those forms of agreements.

10.5. Not sure.

11. I can't conceive of any issues, other than being able to be contacted, which is already an existing issue with licensing, not specific to this. Responsibility for licenses should trickle through the whole process in a chain; the curator should need distribution licenses setting the terms of use for the data set, the developer should abide by the terms of use of the data set and verify the licenses, and the end user should abide by the restrictions imposed on the model and verify the licenses.

12. No. Generative models are black boxes.

13. It'd have a strong chilling effect, no doubt, but the value of such systems is still immense and I believe such systems would still be developed in spite of the increased costs.

14. Nothing I can think of right now. The main thing is that AI models should be considered a distinct use case that is separately licensed from all existing licensing terms, and is not retroactively granted, even by 'wildcard' agreements that permit all uses of the work. Such agreements should only cover uses that are available at the time of licensing.

15. Yes to both. Developers would be unable to meet these requirements if data collectors did not do so as well.

15.1. The amount of works, creator of works, and present owner of rights to works (where it deviates from the creator) should be disclosed, but specific works should not need to be disclosed. However, a list of specific works should be available on request by the creator or rights owner.

15.2. The public and the end users, such that infringement cannot be hidden.

15.3. They should be responsible for disclosures for all models they use; if the models do not come with the appropriate disclosure information, they should not use the model.

15.4. If you were making the records alongside the data collection process, likely hardly anything compared to the licensing costs.

16. Full obligations; if the creator and rights owner have public contact information, attempts should be made and records kept in the case of attached licenses. (think creative commons works) Direct licensing agreements (that is, agreements not originating from the creator and rights owner) should count as filling this obligation.

17. Not sure. I'm not a lawyer.

18. Only when the would-be author has simultaneous licenses to the entire data set behind the model and has meaningfully altered the final product or used the final product as a component of a larger product. I believe in and agree with the interpretation that AI prompts are legally similar to commissioning from an artist, in that the ‘author’ of the work should be the model. (which cannot hold a copyright and therefore cannot produce copyrighted content) The product of a model should have the same usage restrictions as a composite of the most restrictive terms of the constituent licenses.

19. Probably. I don’t know if ‘composite restrictions’ is a thing that is supported by existing laws.

20. No. The incentives that exist right now are more than sufficient.

21. No. A model cannot be an author or inventor when it comes to copyright. We don’t yet have to deal with non-human persons.

22. Yes. If an output can be placed amongst a set of preexisting copyright works and proven to be functionally indistinguishable (a sample of people cannot distinguish the fake from the originals with a statistically significant frequency above mere random chance) it should implicate the rights of the works.

23. Maybe.

24. It should be assumed if the records are not provided or discovered based upon the accessibility of the content to the developer and whether or not the developer employed automated methods of data collection.

25. Depends. If the user tried to trick a model into infringing successfully, it should be the end user. If the developers (of both the model and the wider system) used the content improperly, it should be them. These outcomes should not be mutually exclusive.

25.1. Yes, because technically the end-user should count as a developer in such cases.

26. The developers should be expected to provide a composite set of copyright management information that includes the relevant disclosures and complies with all copyright management information and licenses simultaneously otherwise. Markings should be applied that indicate the origin of the output.

27. I think I made things clear.

28. See 26. The origin of the output (in the form of a brand name, the year of creation, etc.) should be applied and the associated image metadata fields should be populated with similar information.

28.1. The distributor of the AI-generated work in the end, but an honest effort should be made on the part of the service provider.

28.2. Aside from the fact that the distributor/end user can strip labeling/identification with file editing, none that I can think of.

28.3. Violation of all licenses involved in the work’s creation, and the punishments associated with that violation.

29. There are a bunch of them, I can't name any off the top of my head, and the ones presently available are not accurate at all.

30. Impersonation, among other things.

31. Yes, and it should set a floor. Creators have a right to their likenesses, and it should NEVER be possible except through an explicit agreement specifically for this purpose to waive that right. Discrimination should not occur based upon agreement to this explicit agreement's terms or lack thereof, both by the government and by private entities.

32. Yes. The creator should be eligible for these protections, but they should be independent from copyright and specific to AI-generated works.

33. Not sure. Not a lawyer. Definitely will probably need to be clarified with respect to AI though.

34. Can't think of any off the top of my head, and I don't have the time to go looking. This already took me several hours, sorry…